2020

On the Dynamics and Structure of Multiple Strain Epidemic Models and Genotype Networks

Blake Joseph Mitchell Williams
University of Vermont

On the Dynamics and Structure of Multiple Strain Epidemic Models and Genotype Networks

A Thesis Presented

by

Blake Williams

to

The Faculty of the Graduate College

of

The University of Vermont

In Partial Fulfillment of the Requirements
for the Degree of Master of Science
Specializing in Complex Systems and Data Science

October, 2020

Defense Date: September 11, 2020
Thesis Examination Committee:

Laurent Hébert-Dufresne, Ph.D., Advisor
R. Chase Cockrell, Ph.D., Chairperson
Peter Dodds, Ph.D.
Gary An, M.D.
Cynthia J. Forehand, Ph.D., Dean of Graduate College

Abstract

Mathematical disease modeling has long operated under the assumption that any one infectious disease is caused by one transmissible pathogen. This paradigm has been useful in simplifying the biological reality of epidemics and has allowed the modeling community to focus on the complexity of other factors such as contact structure and interventions. However, there is an increasing amount of evidence that the strain diversity of pathogens, and their interplay with the host immune system, can play a large role in shaping the dynamics of epidemics.

This body of work first explores the role of strain-transcending immunity in mathematical disease models, and how genotype networks may be used to explore the evolution of multistrain pathogens. A model is introduced to follow multistrain epidemics with an underlying genotype network. Consequently, the genotype network structure of the antigenic hemagglutinin protein of influenza A (H3N2) is analyzed, suggesting the important role of strain-transcending immunity in the evolution of the virus.

The unique structure of the influenza genotype network is then explored with age-weighted preferential attachment models, utilizing approximate Bayesian computation of the network growth mechanisms. Finally, multistrain vaccination strategies are identified through the application of a genetic algorithm towards minimization of super-critical strains.

Altogether, we show the impact of genotype networks on multistrain disease modeling, explore the role of empirical genotype network structure, and identify applications that include network generative models and vaccine strain selection.

Citations

Material from this thesis has been submitted for publication to PLOS Computational Biology on August 15, 2020, in the following form:

Williams, B.J.M., St-Onge, G., & Hébert-Dufresne, L.. (2020). Localization, unpredictability and epidemic transitions of multistrain epidemics with an underlying genotype network. PLOS Computational Biology.

Pass on what you have learned. Strength, mastery, hmm... but weakness, folly, failure also... we are what they grow beyond.

-Yoda, *Star Wars: The Last Jedi*

Acknowledgements

I thank Dr. Laurent Hébert-Dufresne for embodying all that which an academic advisor should be— an expert researcher, committed to my growth as a scientist, and a role model both within the scientific community and without. I thank all of my labmates and collaborators that I have been fortunate to learn from and work with.

I acknowledge Dr. Maggie Eppstein and Dr. James Bagrow for their role as instructors of courses in which projects have been extended and incorporated into this body of work.

I acknowledge Dr. C. Brandon Ogbunu as the reason I engage in biological research, and by extension Dr. Samuel Scarpino for sparking my interest in infectious diseases and network science.

Finally, I thank all of the members of my committee for generously contributing their time, in support of scientific research and the progression of a young researcher.

Table of Contents

List of Figures

List of Tables

Chapter 0

Introduction

Mathematical disease modeling often operates under the assumption that a single disease is caused by a single pathogen spreading among a population. This paradigm has been useful in simplifying the biological reality of contagions and has allowed the community to focus on the complexity of other factors such as population structure. However, we have long since known of strain diversity within pathogens, and there is an increasing amount of evidence that the diversity of strains and their interplay with the host immune system can play a large role in shaping the dynamics of epidemics [1, 2, 3, 4]. Surveillance efforts have provided unprecedented sequence data for pathogens such as seasonal influenza, enabling a greater understanding of genetic diversity [5]. The increasing availability of genomic data may be used to understand the evolution of the pathogen, and inform epidemic models that incorporate multiple strains and the relationships between them.

The work presented within this document addresses the need to consider mathematical disease models beyond the "one disease, one pathogen" paradigm, regarding: (i) the dynamics within multistrain mathematical epidemic models, (ii) the struc-

ture of influenza A genotype networks and its exploration of genotype space, (iii) an understanding of the generative processes that may determine the structure of viral genotype networks, and (iv) strain-specific considerations for multivalent vaccine selection. We will show that modeling multiple strains is a task that must balance biological realism while maintaining computational tractability. Central to this exploration is the use of the genotype network, a structure that concisely stores strains and the genetic relationships between them. Genotype networks are central to this body of work, with applications in multistrain models, the evolution of influenza A (H3N2), network generative models, and multivalent vaccination strategies.

The following themes are addressed in this thesis. First, we consider the dynamics of epidemics with multiple interacting strains and an underlying genotype network structure (Chapter 1). Particular attention is given towards the endemic state, phase transitions, cyclicity, and localization of infections by strain. This investigation explains the need to consider multiple strains to better understand the dynamics of multistrain pathogens. In Chapter 2 we explore the structure of genotype networks used in the model presented in Chapter 1. Influenza A (H3N2) is used as the model pathogen to leverage the significant quantity of sequence data made available through increased surveillance efforts of the past two decades. The structure of the genotype network is then interpreted in the context of evolution and the effects of cross-protective immunity. A theoretical model of genotype network generation is then explored in Chapter 3 utilizing age-weighted preferential attachment. This exploration relies on approximate Bayesian computation, whose application towards network generative methods is explained more generally. This chapter explores the role of node age within network formation and how it might explain the emergence

of strains within genotype networks. Finally, a method for multivalent vaccine strain selection is developed with consideration for cross-protective immunity in Chapter 4. A genetic algorithm is implemented to provide maximum immune protection across a genotype network under the specified conditions, offering a theoretical vaccination strategy for multiple strains. Together, these themes address the greater need for consideration of multiple strains in mathematical disease modeling.

0.1 Multistrain epidemic models

The mathematical modeling of diseases has been used to influence public health policy, understand the ecology and evolution of pathogens and their hosts, and further the methodology of mathematics, statistics, and more recently, of network science. The construction of an appropriate model can be used to predict disease incidence, determine the effects of climate and animal reservoir population size on risk for human infection, and understand the antigenic variation between strains of a disease. With the careful definition of assumptions, specific processes may be brought to light in an attempt to understand the interaction between humans and pathogens.

The first documented example of mathematical modeling of the spread of disease dates to 1766, when Bernoulli constructed a simple model to justify widespread inoculation against smallpox [6]. This model explored the gains in life expectancy if smallpox were to be eliminated as a cause of death, using differential equations to describe the rate of change for persons who either had not had smallpox or had recovered at a given age [7]. This compartmentalization of persons based on their infection state proved to be an enduring method of organizing a population relative

to some disease.

Early advances on compartmental disease modeling took advantage of the law of mass action, resulting in simple deterministic models. From 1927 to 1933 Kermack and McKendrick developed what is referred to as the SIR and SIRS models, in which a population is divided into suspectible, infectious, or recovered persons, with differential equations governing the rate of flow between compartments [8, 9, 10]. These models have since become the framework of numerous stochastic and deterministic models [11, 12]. An important caveat to these models is their simplicity, with the often fragile assumption of mass action, described as the homogeneous mixing of persons within a population. This implies that each person is equally likely to come into contact with any other, without consideration for contact structure. Varying degrees of correction exist for heterogeneous mixing, such as pair approximations to account for clustering effects, to offer an intermediate level of model complexity between mass action and agent-based models.

The assumptions of mass action models can be improved upon with network epidemiology. A contact network may be introduced to allow for heterogeneity in possible transmission routes, allowing for more realistic spread in human populations than mass action models may allow [13]. Networks can account for heterogeneity at more than one level of disease transmission — we similarly use networks at a different level, that of the viral genotype, to account for different strains of the same pathogen.

Chapter 1 focuses on a deterministic compartmental model of disease dynamics with an underlying genotype network, enabling a straightforward analysis of the endemic state. The endemic state is the number of infections maintained indefinitely in a population under a set of conditions without perturbation. Although disease

transmission is in principle a stochastic process, deterministic modeling provides a versatile framework for studying the endemic state [14]. The effects of demographics on endemic states is in general well understood, but there is a noted need to study the effects of waning immunity (immunity loss at the individual level) on the endemic state, which is of relevance to the cross-protective immunity noted in some pathogens [14, 15].

Accounting for multiple strains within a model may address numerous challenges regarding model realism. For instance, an effective waning immunity may exist as novel strains emerge with different antigenic properties, rendering immunity acquired towards past strains less effective. Allowing for multiple strains within a model allows for a wide variety of modeling choices, which together must balance biological realism with mathematical and computational tractability [16].

Multistrain models are inherently high-dimensional and must take advantage of reductions via symmetry to reduce complexity. A history-based compartmental model with n strains, in which all past infections of a person would be accounted for, contains 2^n combinations of infections a person could have as their history [1]. With the number of differential equations governing this model proportional to 2^n, simplifications must occur to investigate situations in which a practical number of strains are considered.

Of importance to multistrain models and their reduction is the antigenic distance between strains. The human immune system recognizes pathogens by some part of their molecular structure, known as an antigen, to which an antibody may be produced to neutralize the pathogen. The antigenic properties may differ between strains, affecting the body's ability to respond effectively given a novel structure. Antigenic

differences may develop gradually through site mutation, known as antigenic drift, or suddenly such as through genetic reassortment, known as antigenic shift [17]. The former is responsible for continual updates to seasonal influenza vaccines, to keep up with the antigenic features of circulating and not past strains [**?**]. The latter may produce more extreme antigenic changes, which has been responsible for numerous influenza pandemics in the last century [18].

Modeling multistrain pathogens with consideration for antigenic properties requires the inclusion of cross-protective effects, in which the immunity acquired towards one strain offers partial protection towards another strain based on their antigenic similarity. Cross-protection is seen within orthopoxviruses, seasonal influenza subtypes, ebolavirus species, and other pathogens [19, 20, 21]. In general, more similar strains will have greater cross-protective effects, as with influenza A [22]. However, cross-protective immunity is not necessarily a monotonically decreasing function of antigenic distance. Antibody-dependent enhancement has been observed in dengue viruses, in which a past infection may in fact increase the risk of severe infection [23]. Regardless, approximations may be made through antigenic distance. Instead of a unique cross-protective relationship for all combinations of strains, the relationship may be defined as a function of the antigenic distance. This results in antigenic neighborhoods of different antigenic distances from a strain [24]. This simplifying assumption of symmetry reduces model complexity from $\mathcal{O}(2^n)$ to $\mathcal{O}(n*(m+1))$ when up to m neighborhoods of antigenic distance are considered. Given that influenza A will be the pathogen of interest for all chapters, cross-protective immune effects will be modeled by an exponentially decaying function of genetic distance, which may be used as an approximation of antigenic distance [22].

Multistrain models must also consider strain mutation. Although strain mutation is an inherently stochastic process, we choose a deterministic model in this body of work, given the focus on short-term evolutionary trajectory and infection localization in genotype space that may result from cross-protective effects. High strain diversity is to be expected in deterministic models of mutation, given the absence of extinction events removing strains from circulation [25]. As such, the localization of infections within genotype space will be commented on, rather than the presence or absence of particular strains. However, it is possible to incorporate discrete stochastic strain emergence and extinction within the system of equations for a multistrain compartmental model [26]. Regardless of implementation, mutations occur frequently in many pathogens, with RNA viruses such as influenza having significant mutation rates [27, 28]. This necessitates the inclusion of mutation for multistrain models, allowing for evolution from one strain to another. Although mutation allows for the emergence of novel strains, the success of strains will depend on a fitness landscape influenced by numerous factors that include host population immunity.

0.2 Viral genotype networks

Genotype networks are efficient structures used to relate strains through their genetic similarity and plausible evolutionary pathways. A network, also referred to as a graph, consists of nodes that represent objects, and edges connecting nodes if some relationship exists between them. In a genotype network, the nodes are strains defined by their unique genetic sequences, with edges existing between strains whose sequences differ by just one nucleic acid or whose proteins differ by one amino acid,

indicating a plausible mutation pathway [29]. Genotype networks may be considered complementary to phylogenetic trees. Relative to phylogenetic trees, genotype networks allow for cycles indicative of convergent evolution at the cost of having to infer ancestor-descendant directionality.

Selecting an appropriate sequence from the genome of a pathogen, for purposes of defining a strain, will maximize the influence of a mutation on the antigenic properties of the pathogen. Immunoassays are able to concisely describe the antigenic relationship between two strains, namely how well antibodies developed against one strain protects against the another strain. However, immunoassays require all relationships between a set of virus samples to be sampled, which may be unavailable. Instead, identification of a highly antigenic region of a pathogen, as well as epitope regions where an antibody binds to an antigen, can identify genomic regions that are predictive of antigenic properties [30].

A genotype network has previously been defined for influenza A (H3N2) based on the sequence of its surface protein hemagglutinin (HA) [29]. HA is a highly antigenic region of the influenza virion, whose function is to bind the virion to the surface of a host cell and then facilitate the entry of the viral genome for replication within the host cell [31] . HA contains numerous epitopes, making it an important target for the human immune system. As a result, positive Darwinian selection has been observed with ever-changing HA sequences, evolving away from past strains that have influenced the immune profile of the population [2**?**].

In 2014 Wagner constructed a genotype network for HA of influenza A (H3N2), a subtype of seasonal influenza A that has circulated widely since 2010 alongside subtype H1N1 [29]. This network was shown to have cycles that indicated the pres-

ence of multiple mutations at the same site, as well as convergent evolution. The network featured distinct communities and a tree-like structure, with regions of low connectivity spanning the genotype space between high-degree hubs that had numerous genetic neighbors. The presence of strong community structure agrees with that of antigenic clusters found using reduced dimensions of HA inhibition assays and the corresponding phylogenetic trees [32]. The degree of a strain indicated the number of sampled genetic neighbors it had within a distance of one amino acid that could be spanned by a single amino acid substitution. The degree distribution of the network was characteristically heavy-tailed, indicating a high prevalence of strains with few genetic neighbors and a low prevalence of strains with many neighbors. This reflects the trunk-like nature of the phylogenetic trees of seasonal influenza, with periods of strain growth followed by extinction in a forward movement through genotype space, away from past strains [33, 2].

Seasonal influenza surveillance has increased dramatically since 2008, warranting a network analysis of influenza sequences obtained beyond the samples from 2002 to 2007 used by Wagner [29, 34, 35]. Among 35 countries partnering with the CDC, the number of influenza specimens processed per year increased from approximately 100,000 or fewer from 2004 to 2008, to more than 300,000 per year from 2009 to 2013 [34]. Increased surveillance allows for greater coverage of circulating strains, which benefits genotype networks in particular due to the precision required at the level of one amino acid to construct the network. Increased surveillance will produce a less sparse and fragmented genotype network, better capturing the evolutionary pathways between strains.

Genotype networks may be used with multistrain models due to their ability to

capture mutation pathways. Edges within a genotype network indicate a plausible mutation pathway between two strains that may be bridged by just one amino acid substitution [29]. A genotype network consists of the region of genotype space that a pathogen has been observed to explore, providing both known strains and the plausible mutation pathways between them. As such a genotype network provides a set of strains for a multistrain model that may be either model networks or networks informed by real-world observations. Multistrain models informed by genotype networks thus offer a bridge between existing literature on dynamics of multistrain models and the wealth of genomic data available for multistrain pathogens such as seasonal influenza.

0.3 Network generative models

Real-world networks such as genotype networks often fall into distinctive categories based on their topologies, which may be explained through relevant network generative models. Networks are complex structures, with information contained in their topology that may be extracted by understanding the process by which the network formed. Network generative models are tools used to construct networks, explain features of existing networks, and give insight into the specific processes that govern the growth of networks.

The degree distribution of a network alone may be used to broadly classify networks. A commonly used network generative model is the random graph, or Erdős-Rényi network [36]. In this generative model some number of nodes is specified, and all possible edges exists at some common probability independent of one another.

Erdős-Rényi networks have a characteristic binomial degree distribution that converges on Poisson in the limit of many nodes, producing a degree distribution that can be described in full by the mean of a Poisson distribution.

A common generative model with a notably different structure is the small-world network model [37, 38, 39, 40]. In small-world networks, a ring lattice is constructed to produce high local clustering. Edges in the network are then rewired, replacing the end node of some edges with a random node. Depending on the proportion of edges rewired, the final network is often similar, except with a notable reduction in the average shortest path between nodes, hence the name small-world. Small-world networks models are capable of reproducing structure found in social networks and some neuronal networks of the brain [41, 42].

One of the more prominent categories of networks is those that take a heavy-tailed degree distribution, at times similar to a power-law distribution among others [43, 44, 45]. Many real-world networks have a large number of nodes with low degree and a small number of nodes with high degree, such as the internet and electrical power grids [43]. This distribution may be found to remain scale-invariant, and in such a case is known as a scale-free network. The Barabási-Albert model offers an explanation for scale-free behavior with degree-based preferential attachment: a network is grown one node at a time, adding edges to existing nodes in proportion to their degree. The resultant network contains a scale-free degree distribution, in which a power law closely fits the distribution.

If a network is known to have been formed by some generative process, the corresponding model may be able to closely reproduce the network with some parameterization. Since generative processes are typically stochastic, this parameter may follow

a distribution of values capable of producing some observed network. This parameter distribution may be inferred through Bayesian methods, given some observed network and the target. Bayesian inference is a power statistical method, gleaning information on the conditional probability of an event as evidence becomes available.

0.4 Approximate Bayesian computation

In cases where a likelihood function is intractable, approximate Bayesian computation (ABC) may be used to approximate the posterior distribution of a parameter given some prior distribution [46, 47, 48, 49]. A versatile ABC algorithm is rejection sampling, in which parameter values are first drawn from some prior distribution to generate data. The generated data is then compared to the observed data with some summary statistic, and if the difference between summary statistics is below some level of error tolerance, the parameter value is added to the posterior distribution. This posterior distribution is an approximation, converging on the true posterior distribution with reductions in error tolerance.

The parameterization of a generative model may be determined through the use of rejection sampling ABC [50]. A network may be generated by a model under some parameters, and if the network is similar enough in structure to the target network, the parameters will be considered to belong to an approximation of the true posterior distribution. As the tolerance for similarity shrinks towards zero, the computationally determined posterior distribution converges on the true posterior parameter distribution. This requires some summary statistic to compare the generated graph graph to the target. Numerous graph distances exist, such as graph edit distance, NetSimile,

maximum common subgraph, and Laplacian spectral distance, among other spectral distances [51, 52, 53, 54]. Each distance captures the structural differences between graphs in different ways, with various algorithmic complexity: counting the number of nodes and edges that must be added or removed, performing a function on summary statistics such as degree and local clustering, or comparing the difference in eigenvalues of the adjacency or Laplacian matrices of two networks [54]. A distance measure that is highly representative of the difference between between two graphs, or those features desired for comparison, will produce greater confidence in the approximation of the posterior parameter distribution. Appropriate choice of this distance will optimize the ability of ABC to estimate the true posterior distribution.

Identifying the growth mechanisms of genotype networks in particular could help us understand the evolution of the pathogen itself. The preference for nodes by age or degree may indicate which regions of a network contain strains that are actively circulation or where new strains are likely to attach to a network. Such an understanding could influence predictive modeling of strains or even be used incorporated into vaccination strategies.

0.5 Vaccine strain selection

High mutation rates in RNA viruses such as *Zaire ebolavirus*, *Influenza A virus*, and *Rabies lyssavirus* lead to numerous contemporaneous strains [55, 56, 57**?**]. Vaccines are developed based on the antigenic properties of such viruses, however vaccine effectiveness can be less than ideal: influenza vaccine efficacy is commonly below 50% [58, 59, 60, 61, 62]. Effective vaccination is challenged by both rapid evolution

of viruses away from the antigenic properties of strain(s) used for vaccines, and the proper selection of strains for vaccines such that antibodies have a wide-reaching effect on prevalent and future strains [63, 64, 65].

Each spring and fall, the World Heath Organization (WHO) makes recommendations for specific strains to be included in the influenza vaccine for each hemisphere. WHO bases their recommendations largely on the current and forecasted incidence of a particular strain in the upcoming flu season, as well as the availability of similar vaccine viruses [66]. Although attention is given to the genetic and antigenic similarity between strains, further exploitation of available genomic data may be used to inform their recommendations.

Chapter 4 will explore an approximation of vaccine efficacy through suppression of outbreak potential in the presence of vaccinated strains. Cross-protective effects of immunity, observed in viruses such as influenza, allow for genetically similar strains to be influenced by nearby vaccines [67]. A genetic algorithm may be leveraged to find ideal vaccination strains for a given genotype network, given the complexity associated with computation of interacting immune effects. Genetic algorithms evaluate the effectiveness of a solution to a problem according to some measure of fitness. Solutions within a population are recombined, with selection preserving high-performing solutions. This results in the convergence of randomly defined solutions on a high performing if not optimal solution [68]. In the application to follow, this will optimize the location of vaccination strains to maximize immune coverage among observed strains.

Bibliography

[1] Viggo Andreasen, Juan Lin, and Simon A. Levin. The dynamics of cocirculating influenza strains conferring partial cross-immunity. *Journal of Mathematical*

Biology, 35(7):825–842, 1997.

[2] Colin A. Russell, Terry C. Jones, Ian G. Barr, Nancy J. Cox, Rebecca J. Garten, Vicky Gregory, Ian D. Gust, Alan W. Hampson, Alan J. Hay, Aeron C. Hurt, Jan C. de Jong, Anne Kelso, Alexander I. Klimov, Tsutomu Kageyama, Naomi Komadina, Alan S. Lapedes, Yi P. Lin, Ana Mosterin, Masatsugu Obuchi, Takato Odagiri, Albert D. M. E. Osterhaus, Guus F. Rimmelzwaan, Michael W. Shaw, Eugene Skepner, Klaus Stohr, Masato Tashiro, Ron A. M. Fouchier, and Derek J. Smith. The global circulation of seasonal influenza A (H3N2) viruses. *Science*, 320(5874):340–346, 2008.

[3] Thomas Francis. A new type of virus from epidemic influenza. *Science*, 92(2392):405–408, 1940.

[4] Renato Casagrandi, Luca Bolzoni, Simon A. Levin, and Viggo Andreasen. The SIRC model and influenza A. *Mathematical Biosciences*, 200(2):152–169, 2006.

[5] Silvie Van den Hoecke, Judith Verhelst, Marnik Vuylsteke, and Xavier Saelens. Analysis of the genetic diversity of influenza A viruses using next-generation DNA sequencing. *BMC Genomics*, 16(1):79, 2015.

[6] Klaus Dietz and J.A.P. Heesterbeek. Daniel Bernoulli's epidemiological model revisited. *Mathematical Biosciences*, 180(1):1 – 21, 2002.

[7] Sally Blower and Daniel Bernoulli. An attempt at a new analysis of the mortality caused by smallpox and of the advantages of inoculation to prevent it. 1766. *Reviews in Medical Virology*, 14(5):275–288, Sep-Oct 2004.

[8] William Ogilvy Kermack, A. G. McKendrick, and Gilbert Thomas Walker. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 115(772):700–721, 1927.

[9] William Ogilvy Kermack, A. G. McKendrick, and Gilbert Thomas Walker. Contributions to the mathematical theory of epidemics. II. The problem of endemicity. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 138(834):55–83, 1932.

[10] William Ogilvy Kermack, A. G. McKendrick, and Gilbert Thomas Walker. Contributions to the mathematical theory of epidemics. III. Further studies of the problem of endemicity. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 141(843):94–122, 1933.

[11] O. Diekmann and J.A.P. Heesterbeek. *Mathematical epidemiology of infectious diseases: model building, analysis and interpretation*. Wiley series in mathematical and computational biology. John Wiley and Sons, United States, 2000.

[12] Nicholas C. Grassly and Christophe Fraser. Mathematical models of infectious disease transmission. *Nature Reviews Microbiology*, 6(6):477–487, 2008.

[13] Romualdo Pastor-Satorras, Claudio Castellano, Piet Van Mieghem, and Alessan-

dro Vespignani. Epidemic processes in complex networks. *Reviews of Modern Physics*, 87:925–979, Aug 2015.

[14] Mick Roberts, Viggo Andreasen, Alun Lloyd, and Lorenzo Pellis. Nine challenges for deterministic epidemic models. *Epidemics*, 10:49 – 53, 2015.

[15] D Breda, O Diekmann, W F de Graaf, A Pugliese, and R Vermiglio. On the formulation of epidemic models (an appraisal of Kermack and McKendrick). *Journal of Biological Dynamics*, 6 Suppl 2:103–117, 2012.

[16] Adam J. Kucharski, Viggo Andreasen, and Julia R. Gog. Capturing the dynamics of pathogens with many strains. *Journal of Mathematical Biology*, 72(1):1–24, 2016.

[17] John Treanor. Influenza vaccine — outmaneuvering antigenic shift and drift. *New England Journal of Medicine*, 350(3):218–220, 2004.

[18] Edwin D Kilbourne. Influenza pandemics of the 20th century. *Emerging Infectious Diseases*, 12(1):9–14, 01 2006.

[19] Iuliia Gilchuk, Pavlo Gilchuk, Gopal Sapparapu, Rebecca Lampley, Vidisha Singh, Nurgun Kose, David L Blum, Laura J Hughes, Panayampalli S Satheshkumar, Michael B Townsend, Ashley V Kondas, Zachary Reed, Zachary Weiner, Victoria A Olson, Erika Hammarlund, Hans-Peter Raue, Mark K Slifka, James C Slaughter, Barney S Graham, Kathryn M Edwards, Roselyn J Eisenberg, Gary H Cohen, Sebastian Joyce, and Jr Crowe, James E. Cross-neutralizing and protective human antibody specificities to poxvirus infections. *Cell*, 167(3):684–694, 10 2016.

[20] Suzanne L Epstein and Graeme E Price. Cross-protective immunity to influenza A viruses. *Expert Review of Vaccines*, 9(11):1325–1341, 2010.

[21] Lisa E Hensley, Sabue Mulangu, Clement Asiedu, Joshua Johnson, Anna N Honko, Daphne Stanley, Giulia Fabozzi, Stuart T Nichol, Thomas G Ksiazek, Pierre E Rollin, Victoria Wahl-Jensen, Michael Bailey, Peter B Jahrling, Mario Roederer, Richard A Koup, and Nancy J Sullivan. Demonstration of cross-protective vaccine immunity against an emerging pathogenic Ebolavirus species. *PLoS Pathogens*, 6(5):e1000904–e1000904, 05 2010.

[22] Ben Peeters, Sylvia Reemers, Jos Dortmans, Erik de Vries, Mart de Jong, Saskia van de Zande, Peter J M Rottier, and Cornelis A M de Haan. Genetic versus antigenic differences among highly pathogenic H5N1 avian influenza A viruses: consequences for vaccine strain selection. *Virology*, 503:83–93, Mar 2017.

[23] Scott B Halstead. Neutralization and antibody-dependent enhancement of dengue viruses. *Advances in Virus Research*, 60:421–467, 2003.

[24] Neil Ferguson and Viggo Andreasen. The influence of different forms of cross-protective immunity on the population dynamics of antigenically diverse pathogens. In *Mathematical Approaches for Emerging and Reemerging Infectious Diseases: Models, Methods, and Theory*, pages 157–169, New York, NY, 2002.

Springer New York.

[25] Pavlo Minayev and Neil Ferguson. Improving the realism of deterministic multi-strain models: implications for modelling influenza a. *Journal of The Royal Society Interface*, 6(35):509–518, 2009.

[26] Katia Koelle, Meredith Kamradt, and Mercedes Pascual. Understanding the dynamics of rapidly evolving pathogens through modeling the tempo of antigenic change: Influenza as a case study. *Epidemics*, 1(2):129 – 137, 2009.

[27] Rafael Sanjuán, Miguel R. Nebot, Nicola Chirico, Louis M. Mansky, and Robert Belshaw. Viral mutation rates. *Journal of Virology*, 84(19):9733–9748, 2010.

[28] Eri Nobusawa and Katsuhiko Sato. Comparison of the mutation rates of human influenza A and B viruses. *Journal of Virology*, 80(7):3675–3678, 2006.

[29] Andreas Wagner. A genotype network reveals homoplastic cycles of convergent evolution in influenza A (H3N2) haemagglutinin. *Proceedings of the Royal Society B: Biological Sciences*, 281(1786):20132763, 2014.

[30] Jonathan M. Gershoni, Anna Roitburd-Berman, Dror D. Siman-Tov, Natalia Tarnovitski Freund, and Yael Weiss. Epitope mapping. *BioDrugs*, 21(3):145–156, 2007.

[31] D C Wiley and J J Skehel. The structure and function of the hemagglutinin membrane glycoprotein of influenza virus. *Annual Review of Biochemistry*, 56:365–394, 1987.

[32] Derek J. Smith, Alan S. Lapedes, Jan C. de Jong, Theo M. Bestebroer, Guus F. Rimmelzwaan, Albert D. M. E. Osterhaus, and Ron A. M. Fouchier. Mapping the antigenic and genetic evolution of influenza virus. *Science*, 305(5682):371–376, 2004.

[33] Paul G. Thomas and Tomer Hertz. Constrained evolution drives limited influenza diversity. *BMC Biology*, 10(1):43, 2012.

[34] Lauren S Polansky, Sajata Outin-Blenman, and Ann C Moen. Improved global capacity for influenza surveillance. *Emerging Infectious Diseases*, 22(6):993–1001, 06 2016.

[35] Yun Zhang, Brian D Aevermann, Tavis K Anderson, David F Burke, Gwenaelle Dauphin, Zhiping Gu, Sherry He, Sanjeev Kumar, Christopher N Larsen, Alexandra J Lee, Xiaomei Li, Catherine Macken, Colin Mahaffey, Brett E Pickett, Brian Reardon, Thomas Smith, Lucy Stewart, Christian Suloway, Guangyu Sun, Lei Tong, Amy L Vincent, Bryan Walters, Sam Zaremba, Hongtao Zhao, Liwei Zhou, Christian Zmasek, Edward B Klem, and Richard H Scheuermann. Influenza research database: An integrated bioinformatics resource for influenza virus research. *Nucleic Acids Research*, 45(D1):D466–D474, 01 2017.

[36] E. N. Gilbert. Random graphs. *Ann. Math. Statist.*, 30(4):1141–1144, 1959.

[37] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of small-world networks. *Nature*, 393(6684):440–442, 1998.

[38] M.E.J. Newman and D.J. Watts. Renormalization group analysis of the small-world network model. *Physics Letters A*, 263(4):341 – 346, 1999.

[39] M. E. J. Newman and D. J. Watts. Scaling and percolation in the small-world network model. *Physical Review E*, 60:7332–7342, Dec 1999.

[40] A. Barrat and M. Weigt. On the properties of small-world network models. *The European Physical Journal B - Condensed Matter and Complex Systems*, 13(3):547–560, 2000.

[41] L. A. N. Amaral, A. Scala, M. Barthélémy, and H. E. Stanley. Classes of small-world networks. *Proceedings of the National Academy of Sciences*, 97(21):11149–11152, 2000.

[42] Danielle Smith Bassett and Ed Bullmore. Small-world brain networks. *The Neuroscientist*, 12(6):512–523, 2006.

[43] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509, 10 1999.

[44] Albert-László Barabási and Eric Bonabeau. Scale-free networks. *Scientific American*, 288(5):60–69, 2003.

[45] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Physical Review Letters*, 86:3200–3203, Apr 2001.

[46] Mark A. Beaumont. Approximate Bayesian computation in evolution and ecology. *Annual Review of Ecology, Evolution, and Systematics*, 41:379–406, 2010.

[47] Katalin Csilléry, Michael G.B. Blum, Oscar E. Gaggiotti, and Olivier François. Approximate Bayesian computation (ABC) in practice. *Trends in Ecology & Evolution*, 25(7):410 – 418, 2010.

[48] Jarno Lintusaari, Michael U. Gutmann, Ritabrata Dutta, Samuel Kaski, and Jukka Corander. Fundamentals and Recent Developments in Approximate Bayesian Computation. *Systematic Biology*, 66(1):e66–e82, 09 2016.

[49] Jean-Michel Marin, Pierre Pudlo, Christian P. Robert, and Robin J. Ryder. Approximate Bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180, 2012.

[50] Oliver Ratmann, Ole Jørgensen, Trevor Hinkley, Michael Stumpf, Sylvia Richardson, and Carsten Wiuf. Using likelihood-free inference to compare evolutionary dynamics of the protein networks of H. pylori and P. falciparum. *PLoS Computational Biology*, 3(11):e230–e230, 11 2007.

[51] Xinbo Gao, Bing Xiao, Dacheng Tao, and Xuelong Li. A survey of graph edit distance. *Pattern Analysis and Applications*, 13(1):113–129, 2010.

[52] Tina Eliassi-Rad Christos Faloutsos Michele Berlingerio, Danai Koutra. Netsimile: A scalable approach to size-independentnetwork similarity. *ArXiv:1209.2684*, 2012.

[53] Horst Bunke and Kim Shearer. A graph distance metric based on the maximal common subgraph. *Pattern Recognition Letters*, 19(3):255 – 259, 1998.

[54] Peter Wills and François G. Meyer. Metrics for graph comparison: A practitioner's guide. *PLoS ONE*, 15(2):1–54, 02 2020.

[55] Kendra J Alfson, Gabriella Worwa, Ricardo Carrion, and Anthony Griffiths. Spontaneous Mutation Frequency Notes. *Journal of Virology*, 90(5):2345–2355, 2016.

[56] Yi Guan, Dhanasekaran Vijaykrishna, Justin Bahl, Huachen Zhu, Jia Wang, and Gavin J.D. Smith. The emergence of pandemic influenza viruses. *Protein and Cell*, 1(1):9–13, 2010.

[57] Charles Rupprecht, Ivan Kuzmin, and Francois Meslin. Lyssaviruses and rabies: Current conundrums, concerns, contradictions and controversies. *F1000 Research*, 6(0):1–22, 2017.

[58] Edward A. Belongia, Burney A. Kieke, James G. Donahue, Laura A. Coleman, Stephanie A. Irving, Jennifer K. Meece, Mary Vandermause, Stephen Lindstrom, Paul Gargiullo, and David K. Shay. Influenza vaccine effectiveness in Wisconsin during the 2007-08 season: Comparison of interim and final results. *Vaccine*, 29(38):6558–6563, 2011.

[59] Brendan Flannery, Rebecca J Garten Kondor, Jessie R Chung, Manjusha Gaglani, Michael Reis, Richard K Zimmerman, Mary Patricia Nowalk, Michael L Jackson, Lisa A Jackson, Arnold S Monto, Emily T Martin, Edward A Belongia, Huong Q McLean, Sara S Kim, Lenee Blanton, Krista Kniss, Alicia P Budd, Lynnette Brammer, Thomas J Stark, John R Barnes, David E Wentworth, Alicia M Fry, and Manish Patel. Spread of antigenically drifted influenza A(H3N2) viruses and vaccine effectiveness in the United States during the 2018-2019 season. *The Journal of Infectious Diseases*, 30329:1–8, 2019.

[60] Marie R. Griffin, Arnold S. Monto, Edward A. Belongia, John J. Treanor, Qingxia Chen, Jufu Chen, H. Keipp Talbot, Suzanne E. Ohmit, Laura A. Coleman, Gerry Lofthus, Joshua G. Petrie, Jennifer K. Meece, Caroline Breese Hall, John V. Williams, Paul Gargiullo, La Shondra Berman, and David K. Shay. Effectiveness of non-adjuvanted pandemic influenza A vaccines for preventing pandemic influenza acute respiratory illness visits in 4 U.S. communities. *PLoS ONE*, 6(8):4–10, 2011.

[61] John J. Treanor, H. Keipp Talbot, Suzanne E. Ohmit, Laura A. Coleman, Mark G. Thompson, Po Yung Cheng, Joshua G. Petrie, Geraldine Lofthus, Jennifer K. Meece, John V. Williams, Lashondra Berman, Caroline Breese Hall, Arnold S. Monto, Marie R. Griffin, Edward Belongia, and David K. Shay. Effectiveness of seasonal influenza vaccines in the United States during a season with circulation of all three vaccine strains. *Clinical Infectious Diseases*, 55(7):951–959, 2012.

[62] Michael L. Jackson, Jessie R. Chung, Lisa A. Jackson, C. Hallie Phillips, Joyce Benoit, Arnold S. Monto, Emily T. Martin, Edward A. Belongia, Huong Q.

McLean, Manjusha Gaglani, Kempapura Murthy, Richard Zimmerman, Mary P. Nowalk, Alicia M. Fry, and Brendan Flannery. Influenza vaccine effectiveness in the United States during the 2015–2016 season. *New England Journal of Medicine*, 377(6):534–543, 2017.

[63] D. Steinhauer. Rapid Evolution Of RNA Viruses. *Annual Review of Microbiology*, 41(1):409–433, 1987.

[64] F. Carrat and A. Flahault. Influenza vaccine: The challenge of antigenic drift. *Vaccine*, 25(39-40):6852–6862, 2007.

[65] Scott E. Hensley. Challenges of selecting seasonal influenza vaccine strains for humans with diverse pre-exposure histories. *Current Opinion in Virology*, 8:85–89, 2014.

[66] Colin A. Russell, Terry C. Jones, Ian G. Barr, Nancy J. Cox, Rebecca J. Garten, Vicky Gregory, Ian D. Gust, Alan W. Hampson, Alan J. Hay, Aeron C. Hurt, Jan C. de Jong, Anne Kelso, Alexander I. Klimov, Tsutomu Kageyama, Naomi Komadina, Alan S. Lapedes, Yi P. Lin, Ana Mosterin, Masatsugu Obuchi, Takato Odagiri, Albert D.M.E. Osterhaus, Guus F. Rimmelzwaan, Michael W. Shaw, Eugene Skepner, Klaus Stohr, Masato Tashiro, Ron A.M. Fouchier, and Derek J. Smith. Influenza vaccine strain selection and recent studies on the global migration of seasonal influenza viruses. *Vaccine*, 26:D31 – D34, 2008.

[67] Thomas Rowe, Robert A. Abernathy, Jean Hu-Primmer, William W. Thompson, Xiuhua Lu, Wilina Lim, Keiji Fukuda, Nancy J. Cox, and Jacqueline M. Katz. Detection of antibody to avian influenza A (H5N1) virus in human serum by using a combination of serologic assays. *Journal of Clinical Microbiology*, 37(4):937–943, 1999.

[68] Darrell Whitley. A genetic algorithm tutorial. *Statistics and Computing*, 4(2):65–85, 1994.

Chapter 1

On the Emergence of Multistrain Epidemics with an Underlying Genotype Network

Abstract

Mathematical disease modeling has long operated under the assumption that any one infectious disease is caused by one transmissible pathogen spreading among a population. This paradigm has been useful in simplifying the biological reality of epidemics and has allowed the modeling community to focus on the complexity of other factors such as population structure and interventions. However, there is an increasing amount of evidence that the strain diversity of pathogens, and their interplay with the host immune system, can play a large role in shaping the dynamics of epidemics. Here, we introduce a disease model with an underlying genotype network to account

for two important mechanisms. One, the disease can mutate along network pathways as it spreads in a host population. Two, the genotype network allows us to define a genetic distance across strains and therefore to model the transcendence of immunity often observed in real world pathogens. We study the emergence of epidemics in this model, through its epidemic phase transitions, and highlight the role of the genotype network in driving cyclicity of diseases, large scale fluctuations, sequential epidemic transitions, as well as localization around specific strains of the associated pathogen. More generally, our model illustrates the richness of behaviors that are possible even in well-mixed host populations once we consider strain diversity and go beyond the "one disease equals one pathogen" paradigm.

1.1 Introduction

Viral species are known to often undergo rapid evolution. Since the early 20th century, influenza viruses have been described as having marked variability and unpredictable behavior [1]. Subsequent RNA virus studies of the 20th and 21st century have focused on, among others, the *Zaire ebolavirus*, strains of the SARS-CoV species, and HIV-1, all possessing high mutation rates [2]. These frequent mutations contribute to the antigenic evolution of these viruses, allowing them to evade recognition by the human immune system [3].

Despite the long-standing knowledge of subtypes and strains within viral species, mathematical disease modeling has continued to model viral diseases with one underlying pathogen. Notably, influenza violates the "one disease, one pathogen" paradigm: numerous types, subtypes, and strains of influenza viruses challenge the human im-

mune system, driving vaccine effectiveness below 50% in most recent years [4, 5, 6, 7]. Models which fail to account for antigenic variation of a pathogen may lead to biased characterizations of epidemic emergence and progression.

Modeling multistrain pathogens with consideration for antigenic properties requires the inclusion of cross-protective effects, in which the immunity acquired towards one strain offers partial protection towards another strain based on their antigenic similarity. Cross-protection is seen in numerous viral species [8, 9, 10]. In general, more similar strains will have greater cross-protective effects, as with seasonal influenza [11]. However, cross-protective immunity is not necessarily a monotonically decreasing function of antigenic distance. Antibody-dependent enhancement has been observed in dengue viruses, in which a past infection may in fact increase the risk of severe infection [12, 13]. Regardless, approximations of cross-protection may be made through antigenic distance or genetic distance. This relationship may be determined by a function of genetic distance to approximate the unique antigenic distances between all strains.

Several models have been proposed in the growing sub-discipline of multistrain disease modeling [14]. These models balance biological assumptions with computational tractability through reduction via symmetry (e.g. antigenic neighborhoods [15]) age structure [16], and deciding to capture either infection history or immune status [14] among other modeling choices. Cross-protective immunity has been explored in two-strain models [17], multistrain models with a restricted number of antigenic loci and alleles [18], and temporary cross-protective immunity in dengue models [19] capable of producing cyclical and chaos-like infection progression. However, the effects of an underlying genotype network structure — governing viable mutation pathways and

genetic distances between strains — have not been thoroughly explored with multi-strain models. Genotype networks consist of nodes that represent strains, with edges connecting strains that differ by one nucleotide or amino acid in some antigenic region of a gene or protein [20]. Genotype networks are a complementary structure to phylogenetic trees, and are a useful way of representing genetic distance necessary for cross-immunity in multistrain models.

Moreover, the genotype network gives us a proxy through which we can specify potential mutation pathways between strains. Mechanisms for pathogen mutation have previously been included in mathematical models[21, 22], often to consider the emergence of antiviral resistance [23, 24, 25, 26]. Particularly, these models predict the emergence of sequential epidemic transitions — with a first epidemic threshold defining the emergence of macroscopic disease spread and a second marking the emergence of treatment resistant strain [24]. However, such models are often limited to only two pathogen strains as they require specification of the fitness cost associated with resistance. We therefore aim to introduce a more general model, allowing large number of strains to mutate along specific network pathways. While this general model could consider a complex fitness landscape over this genotype network, we focus on the case of neutral evolution and show how the previous results discussed here can all co-exist within a single, fairly simple, model.

We introduce a multistrain Susceptible-Infectious-Recovered-Susceptible (multi-strain SIRS) epidemic model with an underlying genotype network, allowing the disease to evolve along plausible mutation pathways as it spreads in a well-mixed population. We then investigate the effects of genotype network structure on the emergence of an endemic state and on the fitness distribution of strains across the

genotype network. Altogether, our results challenge the typical phenomenology of epidemic models. We find two epidemic transitions: one marking the emergence of an endemic state driven by a subset of strains well localized on the genotype network and one marking a delocalization on the network, where all strains now contribute to the endemic stain. Between these thresholds, we find chaos-like behavior which can be maintained for arbitrarily long times, yielding time series with epidemic cycles featuring large unpredictable fluctuations.

1.2 Model

We study the spread on infectious disease within a well-mixed population for a defined genotype network of the chosen pathogen. Our model is as follows.

The underlying epidemiological dynamics correspond to a simple SIRS model, but where we add a genotype network defined as a set of potential mutations, meaning an infection of strain $i \in [1, N]$ can mutate along the network to a neighbouring strain $j \in \mathcal{N}_i$, where $\mathcal{N}_i$ specifies the set of first network neighbours of strain i. Biologically, this network is defined such that neighbouring strains i and j differ by one unit of genetic distance.

The strains spread within a well-mixed host population. Host individuals are defined as susceptible (S) if they possess no immunity to any strain of a disease, see Fig. 1.1. Susceptible individuals progress to infectious state I_i at transmission rate β for every contact with individuals infectious with strain i, occurring at rate βI_i for every susceptible individual. Note that this basic transmission rate is held constant for all strains, as we focus on neutral evolution as a first approximation.

Individuals in I_i can either: (i) recover at rate γ to state R_i and acquire direct immunity for strain i and partial immunity to strain $j \neq i$; or (ii) become infected with strain I_j via mutation at a rate μ for all strains j in $\mathcal{N}_i$. Individuals in R_i will either: (i) lose immunity and progress to S at rate α, or (ii) become infected with strain $j \neq i$ to which they only possessed partial immunity and progress to I_j at a reduced rate β^*, where β^* is an exponentially decaying function of genetic distance between strains i, j. Specifically:

$$\beta^* \propto 1 - e^{-x_{ij}/\Delta} \tag{1.1}$$

where x_{ij} is the genetic distance between strain i, j (approximated by shortest path of length $x_{ij} = x_{ji}$ between strains i, j in the genotype network) and Δ is the rate of immunity transcendence ($0 < \Delta < \infty$).

Altogether, the dynamics of our model can be followed by the following set of ordinary differential equations (ODEs),

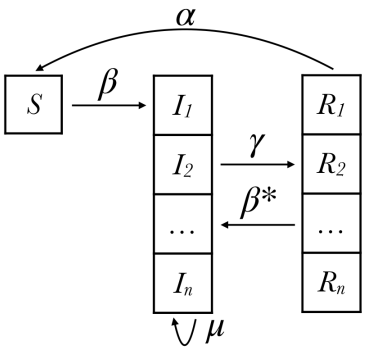
$$\frac{dS}{dt} = -\beta \sum_{i=1}^{N} \frac{SI_i}{N} + \alpha \sum_{i=1}^{N} R_i \tag{1.2}$$

$$\frac{dI_i}{dt} = \beta \frac{SI_i}{N} - \gamma I_i + \mu \sum_{j=1}^{N} A_{i,j}(I_j - I_i) + \beta \sum_{j=1}^{N} \left(1 - e^{-x_{ij}/\Delta}\right) \frac{I_i R_j}{N} \tag{1.3}$$

$$\frac{dR}{dt} = \gamma I_i - \alpha R_i - \beta \sum_{j=1}^{N} \left(1 - e^{-x_{ij}/\Delta}\right) \frac{I_j R_i}{N} \tag{1.4}$$

where A_{ij} is an element of the adjacency matrix of the genotype network, equal to 1 if there is mutation pathway between i and j and 0 otherwise.

We therefore have 5 important epidemiological parameters: transmission rate β, recovery rate γ, rate of waning immunity α, mutation rate μ and immunity transcen-



Symbol	Parameter
n	# strains
N	population size
S	susceptible
I_i	infected (strain i)
R_i	recovered (strain i)
β	transmission rate
γ	recovery rate
α	immunity loss rate
μ	mutation rate
Δ	characteristic length of immunity transcendence
A	adjacency matrix of the genotype network
x_{ij}	shortest path length between strains i & j

Figure 1.1: Compartmental model of a multistrain epidemic with an underlying genotype network **(left)**, with parameters **(right)**.

dence Δ. Unless mentioned otherwise, we typically fix the recovery rate $\gamma = 1$ such that other rates are defined in units of infectious period (and Δ in units of genetic distance). The other parameters then allow us to investigate different interesting regimes,

As $\Delta \to 0$, immunity becomes strain specific with no cross-protective effects. As $\Delta \to \infty$, immunity becomes broad-reaching to the point of universal protection across all strains. With β^* as a function of distance, we are able to reduce model complexity by avoiding specification of β^*_{ij} among all strains, whose values may not be known in real-world applications. Instead, we rely on the inverse relationship between antigenic distance and cross-protection that has been observed in influenza viruses [11]. Note that this relationship may not be monotonically decreasing for all pathogens, in which case β^* may be defined by a function of genetic distance unique to the pathogens.

Our most important assumption is perhaps that only the most recent infection is relevant for cross-immunity effects. Indeed, I_i and R_i specify the pathogen involved

only in the most recent infection for every individual. The alternative would have been to model an infectious state $I_{i,j,\ldots}$ for all unique infection histories, of complexity $\mathcal{O}(2^n)$ if order does not matter and complexity $\mathcal{O}(n!)$ if it does. While a big assumption, focusing on the most recent infection reduces the complexity to $\mathcal{O}(n)$. Computational feasibility would be largely restricted, which would limit the analysis of the effects of genotype network structure [27, 14]. The infection history approximation enables genotype networks to be large enough to contain complex structure, necessary to investigate the role of genotype networks in epidemic progression.

1.3 Results

We focus our attention on the consequences of the genotype network underlying the spread of the disease. In order to gain as much insights as possible on how it affect prevalence of a disease, we keep the network itself simple using well-known graph toy models composed of lattices, chains and stars.

1.3.1 Localization in genotype space

We first ask which strains can be expected to have an advantage, not because of their own fitness or of our epidemiological parameters (as they all share the same β, γ and α), but because of their position in the genotype network. We use three simple network structures — a star, a square lattice, and a chain, all containing 25 strains — and run our model to produce a large outbreak with $\beta = 25$ much greater than the expected SIRS epidemics threshold of $\beta_c = 1$. Accordingly, we set the evolutionary dynamics to be much slower than that of epidemic spread with $\mu = 10^{-3}$. We then

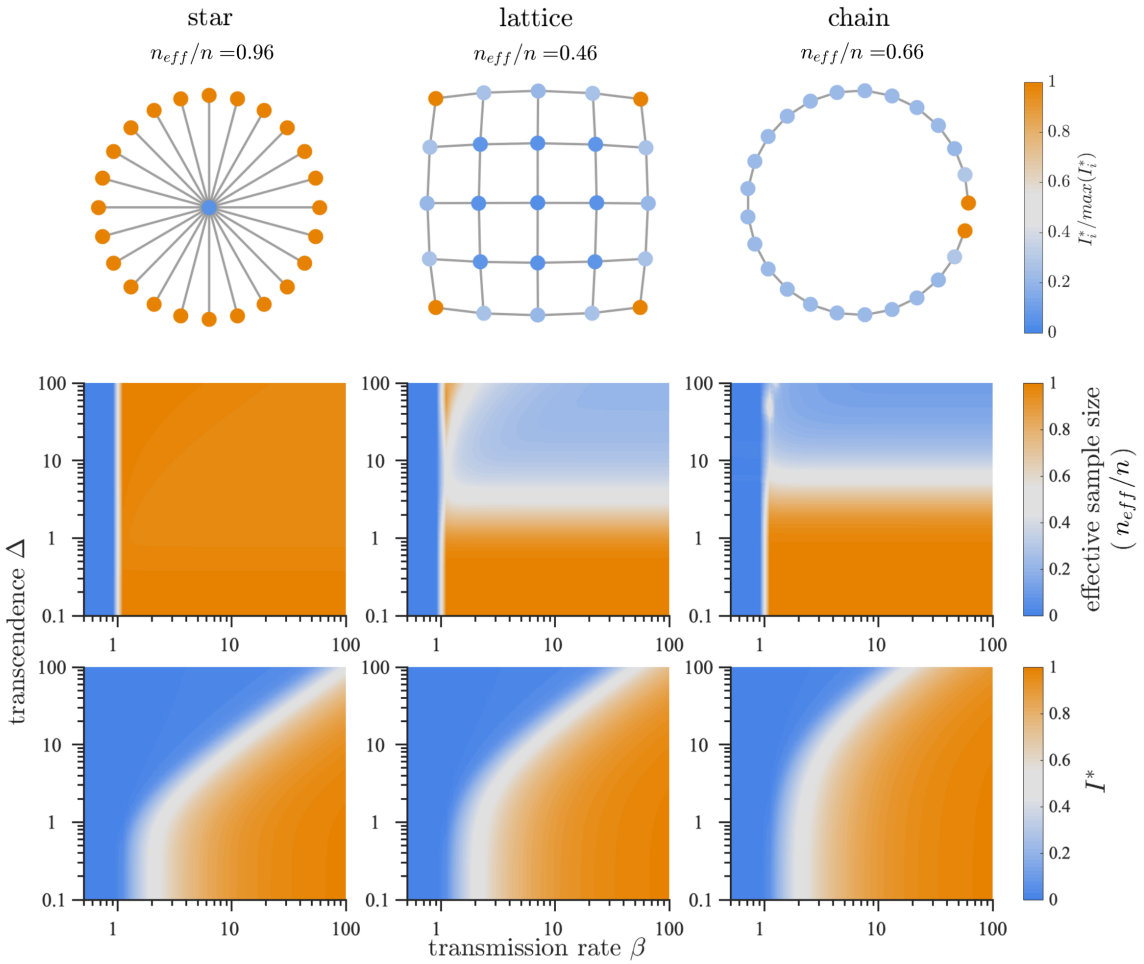


Figure 1.2: Infection localization and characteristics of endemic infection state. **(Top)** Localization within networks shown by endemic infection counts I_i^* normalized by $\max(I_i^*)$ for a given network. We use mutation rate $\mu = 10^{-3}$, transmission rate $\beta = 25$, waning immunity rate $\alpha = 1/50$, and transcending immunity $\Delta = 4$. **(Middle)** Infection localization regimes are revealed where normalized effective sample size is low (lattice and chain), occurring when few strains account for the majority of infections. **(Bottom)** endemic infections depend on not only transmission rate β, but also the breadth of cross-protective effects determined by transcendence rate Δ. Fixed parameters are $n = 25$, $\mu = 10^{-3}$, $\alpha = 1/50$.

let the system reach its endemic steady state, where the derivatives in Eqs. (1.2-1.4) essentially go to zero such that the system is at equilibrium.

We observe a localization of infections by strain within genotype networks as

shown in Fig. 1.2, top row. Stationary or endemic infection counts I_i^* differ from strain to strain, even with the assumption of neutral evolution, based solely on their position in the network and the resulting cross-protective immune effects. Epidemics can therefore be localized around a minority of strains, as is clear in the lattice and chain networks.

We quantify this localization phenomena with Kish's effective sample size [28], referred to here as effective participation ratio $n_{eff}^* = n_{eff}/n = \left(\sum I_i\right)^2 / \left(n \sum I_i^2\right)$. As $n_{eff}^* \to 1$, all strains contribute an equal number of infections. As $n_{eff}^* \to n^{-1}$, only one strain contributes infections. In Fig. 1.2, top row, we observe lower n_{eff}^* in the lattice and chain, indicating greater localization. A small number of strains are able to escape strong cross-protective immunity in the corners of the lattice and at the ends of the chain, while such heterogeneity is not seen in the star and ring networks.

As network structure determines infection localization, so does the transcendence of immunity. In Fig. 1.2, middle row, we see n_{eff}^* as a function of β and immunity transcendence Δ, revealing regimes of strong localization in the lattice and chain networks where n_{eff}^* remains small. High values of $\Delta > 10$, indicating far-reaching cross-protection, are associated with localization in these two networks. The structure of the star and loop networks allow them to escape localization effects influenced by large Δ.

Stationary infection counts I^* are also influenced by immunity transcendence Δ. In Fig. 1.2, bottom row, we see reductions in I^* as Δ increases. As cross-protective effects increase, a higher β becomes necessary to maintain infections. Again we see the importance of network structure, with different values of Δ required between networks to affect I^*.

1.3.2 Sequential phase transitions

We then look at the behaviour of the endemic state as we vary the basic transmission rate β. We know from classic SIRS model that there should be an epidemic transition at $\beta_c = 1$, marking a transition between a disease-free phase where the disease is too weak to establish itself in the population if $\beta < 1$, and an endemic phase for larger values. Yet, one interesting result of Fig. 1.2, bottom row is that the epidemic threshold now seem to increase with transcending immunity Δ. This is somewhat surprising given that Δ does not matter for any one strain, which should still be able to survive on its own following SIRS dynamics once $\beta > \beta_c = 1$.

In Fig. 1.3, we take a deeper look at the phase diagram under varying transmission rate and observe a *second* epidemic transition. More precisely, if α is not too large, I^* is no longer a concave function of the transmission rate; it emerges as expected at $\beta_c = 1$ but has a new inflection point at a much higher β value. This means that only modest increases in I^* are seen when β is just above to the epidemic threshold $\beta_c = 1$, in contrast to standard SIS-like models in which this regime experiences the most rapid rate of change in I^* as a function of β [29].

We conjecture that this second phase transition is governed by what we call the *immune invasion threshold*, corresponding to the point at which infected nodes starts to infect recovered nodes (of other strains) effectively. To see this, in Fig. 1.3(a), we compare the bifurcation diagrams of two models: with and without waning immunity. In the latter case, in the stationary state, a node infected with strain i can only infect recovered nodes of strains $j \neq i$ (since $S^* = 0$). The immune invasion threshold β_{I} can thus be estimated from β_{c} if $\alpha \mapsto 0$. Surprisingly, even though $\beta_{\mathrm{c}} = 1$ whenever

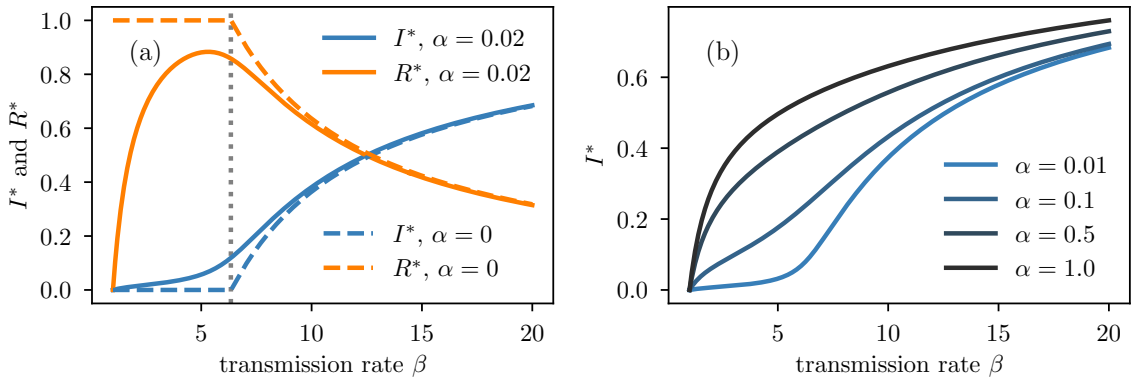


Figure 1.3: Bifurcation diagram for the model with varying levels of waning immunity on a star genotype network with 10 strains. We fix the recovery rate to $\gamma = 1$, the mutation rate $\mu = 1/100$, the transcending immunity $\Delta = 10$ and we vary the transmission rate β. **(a)** We set the waning immunity rate $\alpha \in \{0, 0.02\}$ to illustrate the origin of the immune invasion threshold (vertical dotted line) obtained with Eqs. (1.5) and (1.6). **(b)** For large enough values of waning immunity rate, the immune invasion threshold disappears because recovered nodes quickly become susceptible again.

$\alpha > 0$, it is no longer the case when $\alpha = 0$.

To derive the immune invasion threshold, let us rewrite the stationary state quantities I^*, R^* when $\alpha = 0$. We have

$$0 = -\gamma I_i^* + \mu \sum_j A_{ij} I_j^* - \mu k_i + \beta I_i^* \sum_j T_{ij} R_j^* \,,$$
$$0 = \gamma I_i^* - \beta R_i^* \sum_j T_{ij} I_j^* \,,$$

where $T_{ij} \equiv (1 - e^{-x_{ij}/\Delta})$. Isolating R_i^* in the second equation and reinjecting the solution in the first one, we obtain a self-consistent equation for the $\{I_i^*\}$,

$$I_i^* = \frac{\sum_j A_{ij} I_j^*}{\frac{\gamma}{\mu}\left(1 - \sum_j T_{ij} \frac{I_j^*}{\sum_k T_{jk} I_k^*}\right) + k_i} \,. \tag{1.5}$$

Interestingly, the $\{I_i^*\}$ do not depend upon β. However, we know that such

solution is possible only if $I_i^* > 0 \; \forall i$, and this break down at β_{I} when $R^* = \sum_i R_i^* \to 1$. Therefore, the immune invasion threshold β_{I} has the following explicit expression

$$\beta_{\mathrm{I}} = \gamma \sum_i \frac{I_i^*}{\sum_j T_{ij} I_j^*} \; , \tag{1.6}$$

where the $\{I_i^*\}$ are evaluated from Eq. (1.5).

We observe a direct relationship between Δ and β_{I}. Namely, when $\Delta \to \infty$, $T_{ij} \to 0$ for all i, j, hence $\beta_{\mathrm{I}} \to \infty$, as seen from Eq. (1.6). When $\Delta \to 0$, $T_{ij} \to 1$ for all $i \neq j$, and $T_{ii} = 0$, hence

$$\beta_{\mathrm{I}} \to \gamma \sum_i \frac{I_i^*}{\sum_{j \neq i} I_j *} \; .$$

For large networks, $\beta_{\mathrm{I}} \approx \gamma \equiv 1$ in the limit $\Delta \to 0$. Therefore, we conclude that increasing Δ increases the immune invasion threshold, which makes sense based on intuition alone.

This relationship is shown in Fig. 1.4 for the three toy networks across multiple values of Δ. As Δ increases, immunity becomes wide-reaching in genetic distance, approaching the effects of universal immunity or a universal vaccine. This has the effect of necessitating higher β to produce the same I^* as lower values of Δ. Importantly, because of the sum in the denominator of Eq. (1.6), the immune invasion threshold β_{I} is not simply set by the diameter of the genotype network (i.e., the maximum value of x_{ij}), and is instead set by the entire network structure. While strains maximally distant from each other can of course better infect recovered individuals, competition between strains also play an important role: central strains can still infect individuals and grant them better immunity due to their central position in the network.

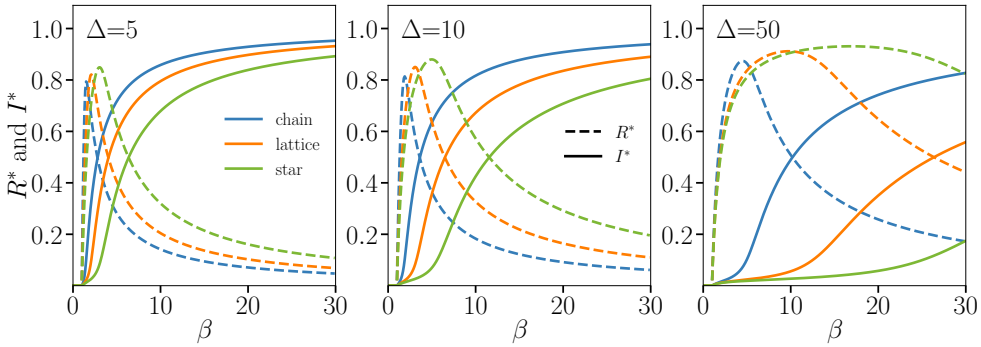


Figure 1.4: Integration of the ODEs on three toy genotype networks — the chain, square lattice and star — all with 25 strains. We fix the recovery rate $\gamma = 1$, the mutation rate $\mu = 1/100$, the waning immunity rate $\alpha = 1/50$ and vary the transmission rate β under three values of transcending immunity: (a) $\Delta = 5$, (b) $\Delta = 10$, (c) $\Delta = 50$. Close to the inflection point of every I^* curve (shown in solid lines) we find a maximum in R^* (shown in dashed lines). This point therefore marks a second activation threshold, one where the transmission rate is high enough to counteract transcending immunity and spread the outbreak using the pool of recovered individuals.

Thus, the network structure plays a nontrivial role in setting the exact value of β_{I} as determined by Eq. (1.6).

1.3.3 Rich dynamics between epidemic thresholds

Beyond the features of the endemic state, we observe rich prevalence dynamics throughout the epidemic when transmission rates are between the epidemic threshold $\beta_c = 1$ and the immune invasion threshold $\beta_{\mathrm{I}} \geq \beta_c$. By comparing the top, middle, and bottom rows of Fig. 1.5 we see infection counts throughout the epidemic simulation while the transmission rate lays in different regimes, decreasing from $\beta > \beta_{\mathrm{I}}$ to values closer to $\beta_c = 1$.

For transmission rate below the immune invasion threshold (bottom two rows), we see oscillations in the overall infection counts across all three networks before

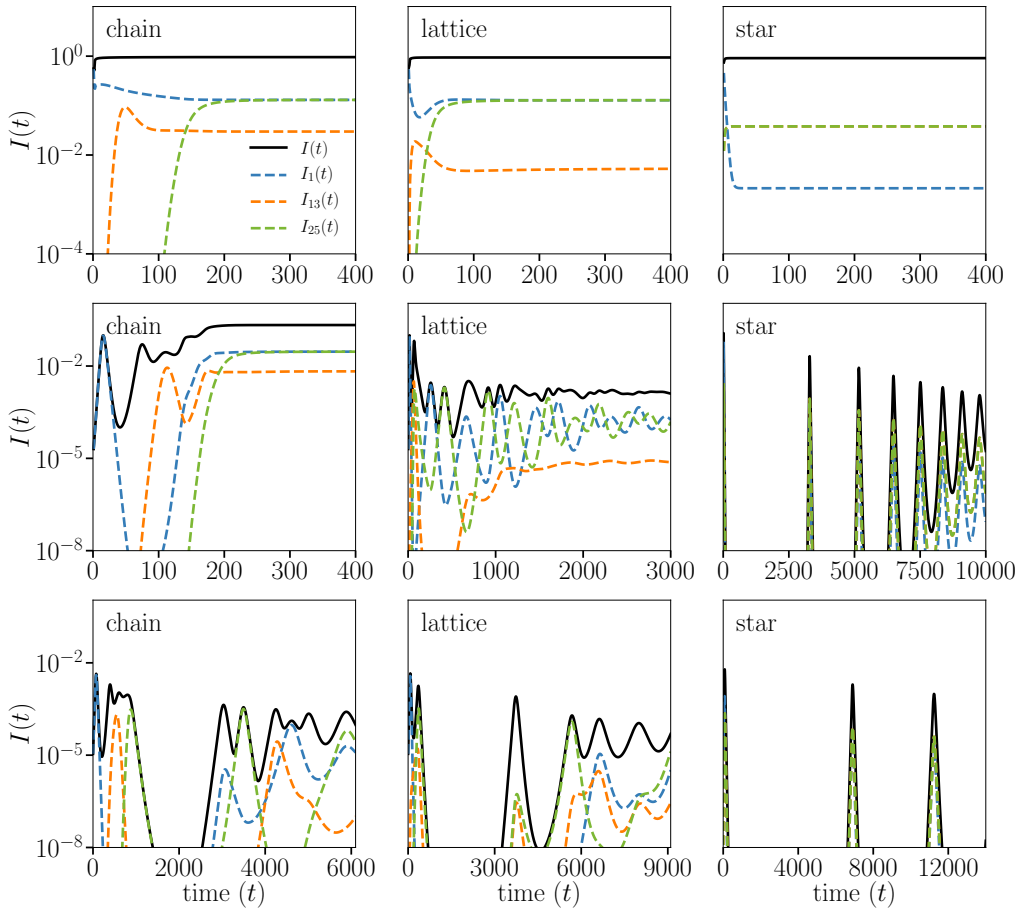


Figure 1.5: Integration of the ODEs on three toy genotype networks — the chain (left column), square lattice (middle column) and star (right column) — all with 25 strains. We fix the recovery rate $\gamma = 1$, the mutation rate $\mu = 1/1000$, the waning immunity rate $\alpha = 1/5000$ and the transcending immunity $\Delta = 4$ and vary the transmission rate: $\beta = 25$ (**top**), $\beta = 1.7$ (**middle**), $\beta = 1.1$ (**bottom**). The system is initialized with a small fraction 10^{-5} of infections on an "end strain" (end for the chain, corner for the lattice, leaf for the star). On the chain we see successive activation of all strains, with the system stabilizing once the entire network is explored and evolution reaches a dead-end. The star sees cycles caused by activation of the leaf strains. The lattice is much more interesting, with loops causing a random-like succession of strains to cycle. The dynamics become more interesting for the bottom row, with transmission rates between the epidemic and immune invasion thresholds, with cycles and chaos-like dynamics. The closer we get to the true epidemic thresholds $\beta_c = 1$, the longer the interesting transient dynamics.

converging on an endemic value, resembling a dampened pseudo-chaotic behavior. Noting the different time scales shown, the chain rapidly converges on its endemic state while the star undergoes drastic oscillations before convergence. We see variation in infection counts at the strain level, with the infection counts for 3 of the 25 strains shown. At the strain level we see convergence occurring on different time scales within the same network, as well as variability in oscillatory nature.

In comparison, the top row of Fig. 1.5 shows the rapid convergence on the endemic state when the transmission rate is high ($\beta = 25$). There still exists infection localization, as indicated by different endemic infection counts at the strain level, as well as variability in convergence time between strains. However, the oscillatory nature is profoundly absent at transmission rates well about β_{I}. In contrast, as transmission rates are lowered towards $\beta_c = 1$ in the bottom rows of Fig. 1.5, we see the oscillations preserved but stretched across a broader timescale. Importantly, as the timescale of oscillations is stretched, their minimal values decrease by orders of magnitude. In practice, this shows that any finite size simulations of the dynamics captured by our model would likely lead to strain extinction, with potential to reemerge through mutations. Discrete events are unfortunately not captured in ODE models as they assume continuous values, or infinite population.

1.4 Conclusion

The introduction of an underlying genotype network to a multistrain model has demonstrated the emergence of cyclicity, infection localization, and sequential phase transitions, all in one model. Simple mathematical arguments have allowed us to

solve for the transitions observed and highlight the nontrivial impact of the structure of the genotype network. Rich infection dynamics are seen between the epidemic threshold and the immune invasion threshold. Altogether, what these results show is that many features of infectious disease dynamics often explained by environmental factors or host behaviour, such as cyclicity [30], unpredictability [31] and sequential transitions [32], can also be explained by adding a layer of biological complexity in the form of a genotype network. Our results thus highlight the importance of going beyond the "one disease, one pathogen" paradigm, with complex dynamics emerging from even the most simple genotype network structures.

Future work needs to be done to integrate this modelling approach with real genomic data. Likewise, the interplay of our results with the finite size and the contact structure of the host population needs to be investigated; as does the role of strain extinction and emergence. Different modelling approaches will need to be considered, such as explicitly modeling the growth and evolution of the genotype network as it co-evolves (albeit on a different timescale) with the spread of the infectious disease in the host population. Coupling the large modeling literature on growing networks [33] with that of network epidemiology [29] should lead to a richer understanding of how networks, both biological and social, impact epidemics. Finally, this type of models could also be appropriate to re-imagine vaccination strategies. The literature on targeted immunization and influential spreaders on networks could then be leveraged [34, 35, 36, 37], but rather than targeting central individuals the objective would be to best hinder and block the immune evasion of the pathogen as it mutates along its genotype network.

In terms of applying these models to specific scenarios, there is a need for unbiased

pathogen genomic data, as well as an understanding of their antigenic properties, to inform models that account for these features using real-world data and to refine the cross-protective immune effects between strains of a pathogen. Similarly, we need more realistic models to take advantage of the growing body of genomic data available and refine the mechanisms driving mutation and immunity. We call for the refinement of immune mechanisms and immune history to allow their incorporation in mathematical disease models. Further understanding of how pathogens explore genotype space, the growth of genotype networks, the role of host immunity towards past strains, and the influence of the above on the fitness landscape of pathogens will better inform models incorporating multiple strains, cross-protective effects, and evolution of a pathogen.

Bibliography

[1] G Meiklejohn and H B Bruyn. Influenza in California during 1947 and 1948. *American Journal of Public Health and the Nation's Health*, 39(1):44–49, 01 1949.

[2] Rafael Sanjuán, Miguel R. Nebot, Nicola Chirico, Louis M. Mansky, and Robert Belshaw. Viral mutation rates. *Journal of Virology*, 84(19):9733–9748, 2010.

[3] J. C. de Jong, D. J. Smith, A. S. Lapedes, I. Donatelli, L. Campitelli, G. Barigazzi, K. Van Reeth, T. C. Jones, G. F. Rimmelzwaan, A. D. M. E. Osterhaus, and R. A. M. Fouchier. Antigenic and genetic evolution of swine influenza A (H3N2) viruses in Europe. *Journal of Virology*, 81(8):4315–4322, 2007.

[4] Brendan Flannery, Jessie Clippard, Richard K Zimmerman, Mary Patricia Nowalk, Michael L Jackson, Lisa A Jackson, Arnold S Monto, Joshua G Petrie, Huong Q McLean, Edward A Belongia, Manjusha Gaglani, LaShondra Berman, Angie Foust, Wendy Sessions, Swathi N Thaker, Sarah Spencer, Alicia M Fry, Centers for Disease Control, and Prevention. Early estimates of seasonal influenza vaccine effectiveness - United States, January 2015. *MMWR. Morbidity and Mortality Weekly Report*, 64(1):10–15, 01 2015.

[5] Brendan Flannery, Jessie R Chung, Swathi N Thaker, Arnold S Monto, Emily T Martin, Edward A Belongia, Huong Q McLean, Manjusha Gaglani, Kempapura

Murthy, Richard K Zimmerman, Mary Patricia Nowalk, Michael L Jackson, Lisa A Jackson, Angie Foust, Wendy Sessions, LaShondra Berman, Sarah Spencer, and Alicia M Fry. Interim estimates of 2016-17 seasonal influenza vaccine effectiveness - United States, February 2017. *MMWR Morbidity and Mortality Weekly Report*, 66(6):167–171, Feb 2017.

[6] Brendan Flannery, Jessie R Chung, Edward A Belongia, Huong Q McLean, Manjusha Gaglani, Kempapura Murthy, Richard K Zimmerman, Mary Patricia Nowalk, Michael L Jackson, Lisa A Jackson, Arnold S Monto, Emily T Martin, Angie Foust, Wendy Sessions, LaShondra Berman, John R Barnes, Sarah Spencer, and Alicia M Fry. Interim estimates of 2017-18 seasonal influenza vaccine effectiveness - United States, February 2018. *MMWR Morbidity and Mortality Weekly Report*, 67(6):180–185, Feb 2018.

[7] Joshua D Doyle, Jessie R Chung, Sara S Kim, Manjusha Gaglani, Chandni Raiyani, Richard K Zimmerman, Mary Patricia Nowalk, Michael L Jackson, Lisa A Jackson, Arnold S Monto, Emily T Martin, Edward A Belongia, Huong Q McLean, Angie Foust, Wendy Sessions, LaShondra Berman, Rebecca J Garten, John R Barnes, David E Wentworth, Alicia M Fry, Manish M Patel, and Brendan Flannery. Interim estimates of 2018-19 seasonal influenza vaccine effectiveness - United States, February 2019. *MMWR Morb Mortal Wkly Rep*, 68(6):135–139, Feb 2019.

[8] Iuliia Gilchuk, Pavlo Gilchuk, Gopal Sapparapu, Rebecca Lampley, Vidisha Singh, Nurgun Kose, David L Blum, Laura J Hughes, Panayampalli S Satheshkumar, Michael B Townsend, Ashley V Kondas, Zachary Reed, Zachary Weiner, Victoria A Olson, Erika Hammarlund, Hans-Peter Raue, Mark K Slifka, James C Slaughter, Barney S Graham, Kathryn M Edwards, Roselyn J Eisenberg, Gary H Cohen, Sebastian Joyce, and Jr Crowe, James E. Cross-neutralizing and protective human antibody specificities to poxvirus infections. *Cell*, 167(3):684–694, 10 2016.

[9] Lisa E Hensley, Sabue Mulangu, Clement Asiedu, Joshua Johnson, Anna N Honko, Daphne Stanley, Giulia Fabozzi, Stuart T Nichol, Thomas G Ksiazek, Pierre E Rollin, Victoria Wahl-Jensen, Michael Bailey, Peter B Jahrling, Mario Roederer, Richard A Koup, and Nancy J Sullivan. Demonstration of cross-protective vaccine immunity against an emerging pathogenic Ebolavirus species. *PLoS Pathogens*, 6(5):e1000904–e1000904, 05 2010.

[10] Suzanne L Epstein and Graeme E Price. Cross-protective immunity to influenza A viruses. *Expert Review of Vaccines*, 9(11):1325–1341, 2010.

[11] Ben Peeters, Sylvia Reemers, Jos Dortmans, Erik de Vries, Mart de Jong, Saskia van de Zande, Peter J M Rottier, and Cornelis A M de Haan. Genetic versus antigenic differences among highly pathogenic H5N1 avian influenza A viruses: consequences for vaccine strain selection. *Virology*, 503:83–93, Mar 2017.

[12] Scott B Halstead. Neutralization and antibody-dependent enhancement of dengue viruses. *Advances in Virus Research*, 60:421–467, 2003.

[13] Leah C. Katzelnick, Lionel Gresh, M. Elizabeth Halloran, Juan Carlos Mercado, Guillermina Kuan, Aubree Gordon, Angel Balmaseda, and Eva Harris. Antibody-dependent enhancement of severe dengue disease in humans. *Science*, 358(6365):929–932, 2017.

[14] J.R. Gog A. J. Kucharski, V. Andreasen. Capturing the dynamics of pathogens with many strains. *Journal of Mathematical Biology*, 72(1-2):1–24, 2016.

[15] Neil Ferguson and Viggo Andreasen. The influence of different forms of cross-protective immunity on the population dynamics of antigenically diverse pathogens. In *Mathematical Approaches for Emerging and Reemerging Infectious Diseases: Models, Methods, and Theory*, pages 157–169, New York, NY, 2002. Springer New York.

[16] A. J. Kucharski and J. R. Gog. Age profile of immunity to influenza: effect of original antigenic sin. *Theoretical Population Biology*, 81(2):102–112, Mar 2012.

[17] Masashi Kamo and Akira Sasaki. The effect of cross-immunity and seasonal forcing in a multi-strain epidemic model. *Physica D: Nonlinear Phenomena*, 165(3):228 – 241, 2002.

[18] Pavlo Minayev and Neil Ferguson. Improving the realism of deterministic multi-strain models: implications for modelling influenza a. *Journal of The Royal Society Interface*, 6(35):509–518, 2009.

[19] Zhilan Feng and Jorge X Velasco-Hernández. Competitive exclusion in a vector-host model for the dengue fever. *Journal of Mathematical Biology*, 35(5):523–544, 1997.

[20] Andreas Wagner. A genotype network reveals homoplastic cycles of convergent evolution in influenza A (H3N2) haemagglutinin. *Proceedings of the Royal Society B: Biological Sciences*, 281(1786):20132763, 2014.

[21] Michelle Girvan, Duncan S Callaway, Mark EJ Newman, and Steven H Strogatz. Simple model of epidemics with pathogen mutation. *Physical Review E*, 65(3):031915, 2002.

[22] Florian Uekermann and Kim Sneppen. Spreading of multiple epidemics with cross immunization. *Physical Review E*, 86(3):036108, 2012.

[23] Marc Lipsitch, Ted Cohen, Megan Murray, and Bruce R Levin. Antiviral resistance and the control of pandemic influenza. *PLoS Med*, 4(1):e15, 2007.

[24] Laurent Hébert-Dufresne, Oscar Patterson-Lomba, Georg M Goerg, and Benjamin M Althouse. Pathogen mutation modeled by competition between site and bond percolation. *Physical Review Letters*, 110(10):108103, 2013.

[25] Benjamin M Althouse, Oscar Patterson-Lomba, Georg M Goerg, and Laurent Hébert-Dufresne. The timing and targeting of treatment in influenza pandemics influences the emergence of resistance in structured populations. *PLOS Comp.*

Bio., 9(2):e1002912, 2013.

[26] Oscar Patterson-Lomba, Benjamin M Althouse, Georg M Goerg, and Laurent Hébert-Dufresne. Optimizing treatment regimes to hinder antiviral resistance in influenza across time scales. *PloS one*, 8(3):e59529, 2013.

[27] M Gabriela M Gomes, Graham F Medley, and D James Nokes. On the determinants of population structure in antigenically diverse pathogens. *Proceedings of the Royal Society B: Biological Sciences*, 269(1488):227–233, 02 2002.

[28] L. Kish. *Survey Sampling*. John Wiley & Sons, New York, 1965.

[29] Romualdo Pastor-Satorras, Claudio Castellano, Piet Van Mieghem, and Alessandro Vespignani. Epidemic processes in complex networks. *Reviews of Modern Physics*, 87:925–979, Aug 2015.

[30] Benjamin M Althouse and Laurent Hébert-Dufresne. Epidemic cycles driven by host behaviour. *Journal of The Royal Society Interface*, 11(99):20140575, 2014.

[31] Samuel V Scarpino and Giovanni Petri. On the predictability of infectious disease outbreaks. *Nature Communications*, 10(1):1–8, 2019.

[32] Antoine Allard, Benjamin M Althouse, Samuel V Scarpino, and Laurent Hébert-Dufresne. Asymmetric percolation drives a double transition in sexual contact networks. *Proceedings of the National Academy of Sciences*, 114(34):8969–8973, 2017.

[33] Sergey N Dorogovtsev and Jose F.F. Mendes. Evolution of networks. *Advances in Physics*, 51(4):1079–1187, 2002.

[34] Romualdo Pastor-Satorras and Alessandro Vespignani. Immunization of complex networks. *Physical Review E*, 65(3):036104, 2002.

[35] Reuven Cohen, Shlomo Havlin, and Daniel Ben-Avraham. Efficient immunization strategies for computer networks and populations. *Physical Review Letters*, 91(24):247901, 2003.

[36] Laurent Hébert-Dufresne, Antoine Allard, Jean-Gabriel Young, and Louis J Dubé. Global efficiency of local immunization on complex networks. *Scientific Reports*, 3:2171, 2013.

[37] Flaviano Morone and Hernán A Makse. Influence maximization in complex networks through optimal percolation. *Nature*, 524(7563):65–68, 2015.

Chapter 2

On the Genotype Network of Influenza A (H3N2) Hemagglutinin

Abstract

Seasonal influenza is a virus of global public health concern, with multiple ever-changing strains in circulation. A high mutation rate enables the influenza virus to evade recognition by the human immune system, including immunity acquired through past infection and vaccination. Here, we capture the genetic similarity of influenza strains and their evolutionary pathways with genotype networks. We show that influenza A (H3N2) hemagglutinin genotype networks are characterized by heavy-tailed component size and degree distributions suggesting critical-like behavior, as well as a growth process suggesting temporally-restricted preferential attachment. We argue that: (i) genotype networks are driven by mutation and host immunity to explore a subspace of networks predictable in structure, and (ii) genotype networks provide an underlying structure necessary to capture the rich dynamics of multi-

strain epidemic models. In particular, inclusion of strain-transcending immunity in epidemic models is dependent upon the structure of an underlying genotype network, revealing edge densities that maximize endemic infections. We conclude that genotype networks provide a model that enables network analysis of pathogen evolution and realistic multistrain epidemic models.

2.1 Introduction

Each year, seasonal influenza results in 290,000 to 650,000 deaths globally, 9 million to 36 million cases in the United States alone, and a significant economic burden [1, 2, 3]. Despite widespread vaccination and increased surveillance efforts in recent years, influenza continues to show prominent seasonality in temperate regions and causes a year-round burden in tropical regions [4]. Risk for severe outcomes is non-trivial in at-risk populations, including young children, the elderly, pregnant women, and persons with certain pre-existing conditions [5]. The risk for severe outcomes in these groups is further elevated in low- and middle-income countries, a concerning relationship given the global prevalence of influenza [6].

Influenza viruses mutate at a high rate, leading to frequent strain emergence [7]. These novel strains may be antigenically different enough to escape recognition in the host by previously acquired antibodies. As a result, circulating strains of influenza are constantly changing, challenging the human immune response and necessitating yearly updates to vaccine strains. Optimal vaccine strain selection is dependent upon the ability to both forecast prevalent future strains and select a limited number of vaccine strains, such that these strains offer optimal immune protection by leveraging strain-

transcending immunity [8, 9]. Modern seasonal influenza vaccines induce antibodies for three to four unique strains of influenza, providing direct immunity for these strains and cross-protective (or strain-stranscending) effects towards antigenically similar strains. Likewise, these antibodies are induced in response to an influenza infection. Antibodies will not offer protection against all strains, but only those identical to or antigenically similar to the strain that induced the antibodies [10]. This finite strain-transcendence of immunity, in which antibodies for one strain offer protection against only antigenically similar strains, is what allows influenza to persist year after year.

Genotype networks have previously been constructed from the highly antigenic hemagglutinin (HA) protein sequences of influenza [11]. The networks revealed features not well represented in phylogenetic trees, such as convergent evolution, the identical trait evolution in separate lineages. These networks were prone to fragmentation in the presence of low sampling rates, reducing the number of observed plausible evolutionary pathways. Sample rates have increased dramatically since Wagner 2014, calling for a more accurate account of the evolution of influenza genotype networks [11].

In this investigation we utilize a genotype network to capture the genetic relationship between strains of influenza A (H3N2), a prominent subtype from 2010 to 2020. Sequences of the highly antigenic HA protein of influenza A (H3N2) will be used to explore the structure and temporal evolution of the genotype network and its exploration of genotype space. Finally, a multistrain epidemic model is implemented to explore the role of edge density in determining infections in the context of strain-transcending immunity.

2.2 Methods

2.2.1 Network generation

Protein sequences were obtained for complete influenza A (H3N2) HA samples from the Influenza Research Database [12]. Samples acquired from the Influenza Research Database are sourced from the databases that include NCBI GenBank and RefSeq. Samples were restricted to a collection date of January 4, 1999 to October 1, 2019 and collection from human hosts only, obtained on January 16, 2020. A 3 month delay between final sample collection date and data retrieval date was implemented to account for delays in data reporting.

A total of 30,175 sequenced samples for HA were obtained. Sequences were further restricted to allow for the precise genetic sequence comparison required for network edge construction. Samples with missing or uncertain residues ($n = 1,278$) and sequences with more or less than 566 amino acids ($n = 17$) were removed. The remaining 28,880 samples were condensed into set V of 6,494 unique sequences.

The number of differing amino acids across all sites, $d_{v,w}$, was found for all pairs of sequences of length l, where $l = 566$:

$$d_{v,w} = \sum_{i=1}^{l} x, \text{where } x = \begin{cases} 1, & \text{if } v_i \neq w_i \\ 0, & \text{if } v_i = w_i \end{cases} \qquad v, w \in V$$

An edge $e_{v,w}$ is formed where $d_{v,w} = 1$. Each edge indicates a plausible, but not definitive, mutation pathway between two viable strains that requires one point mutation, thus no intermediate strains nor multi-mutation events. The resulting genotype net-

work is defined as $G = (V, E)$, where E is the set of all edges $e_{v,w}$.

Temporal analyses restricted data by year using seasonal trends of the Northern Hemisphere, given its dominance of the data set. Sequences were binned according to a 5 year window, where each year consisted of July 1 through June 30 of the following year. A 5 year window centered on 2010 would contain sequences from July 2007 through June 2012.

Distribution tails were fit with power laws using the 'poweRlaw' package [13]. Minimum x values fit were constrained to $x_{min} = 2$ for networks consisting of 5 years of data, while the network across all years was fit using a minimum x value of at least 5, found to be $x_{min} = 7$ for both degree and component size by poweRlaw.

2.2.2 Multistrain epidemic model

The multistrain epidemic model implements the compartment model of Chapter 1 (see Chapter 1, Figure 1.1). Susceptible individuals S may become infected by any strain i at rate β, progressing to infected state I_i. Individuals in I_i either recover to the corresponding recovered state R_i or become infected with a neighboring strain j in the genotype network at mutation rate μ. Recovered individuals in state R_i return to S at immunity loss rate α, or become infected by strain j where $j \neq i$. Transmission from strain j to an individual in R_i is weighted by a decaying function of the distance between strains in the genotype network:

$$\beta^* = \beta(1 - e^{-x_{ij}\Delta^{-1}}) \tag{2.1}$$

for shortest path x_{ij} between strains i, j, and transcendence of immunity parameter Δ.

Immunity remains strain-specific where $\Delta = 0$. Small values of Δ correspond to acquired immunity that protects best against only the most similar strains, with minimal protection against distant strains. Large values of Δ provide partial immune protection of greater range and strength. Note that the shortest path between nodes is an approximation for number of differing amino acids, which may differ slightly.

This compartmental model was implemented with a system of differential equations containing one susceptible state and an infected and recovered state for each strain. The model assumes that: (i) an individual may be infected by at most one strain at a time, (ii) an individual may possess acquired immunity for at most one strain at a time, and (iii) transcendence of immunity decays exponentially as a function of the distance between strains (Equation 2.1).

2.3 Results

2.3.1 Influenza A (H3N2) HA genotype network

The influenza A (H3N2) HA genotype network represents 28,880 samples of HA from 1999 to 2019, resulting in 9,714 nodes (unique strains), 7,599 edges (possible point mutations between strains), and 3,262 connected components, of which 384 consist of more than one node. With 29.6% of nodes of degree $k = 0$ and 44.0% of $k = 1$, the network features a heavy-tailed degree distribution, stretching up to a maximum degree of 256. The tail of the complementary cumulative distribution function (CCDF) of

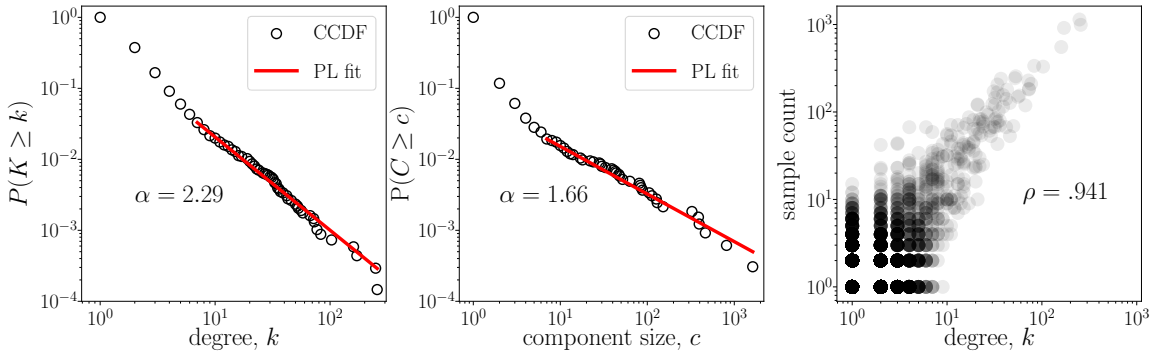


Figure 2.1: Influenza A (H3N2) HA genotype network degree and component size distribution. **(left)** CCDF of degree. The tail of degree distribution does not significantly differ from a power-law distribution with $\alpha_k = 2.29$ for $k_{\min} = 7$ ($p = 0.11$, $\alpha_{\text{significance}} = 0.05$, 10^3 repetition Kolmogorov-Smirnov test). **(center)** CCDF of component size. Component size distribution does not significantly differ from a power-law distribution with $\alpha_c = 1.66$ for $c_{\min} = 7$ ($p = 0.59$, $\alpha_{\text{significance}} = 0.05$, 10^3 repetition Kolmogorov-Smirnov test). **(right)** Sample count of a sequence vs. degree k of corresponding node. Sample count is highly correlated with node degree ($\rho = 0.941$).

degree, $P(K \geq k)$, exhibits power-law behavior: $P(K \geq k) \propto k^{-\alpha_k}$ where $\alpha_k = 2.29$ for minimum degree $k_{min} = 7$ (Figure 2.1, left). This is in agreement with the the heavy-tailed degree distribution found by Wagner in the largest connected component, or the giant component, of a smaller data set from 2002-2007 [11]. This degree distribution suggests that in a generative model of such a network, approximately linear preferential attachment would be necessary to reproduce a comparable degree distribution [14].

The distribution of component sizes of the genotype network is similarly heavy-tailed. The tail of the CCDF of component sizes $P(C \geq c)$ follows a power-law distribution, where $P(C \geq c) \propto c^{-\alpha_c}$ with $\alpha_c = 1.66$ for minimum component size $c_{min} = 7$ (Figure 2.1, center). This scaling may be suggestive of a self-organized critical process in the formation of the genotype network.

The degree of a node and the number of times its corresponding sequence was

sampled are highly correlated (Figure 2.1, right). This suggests that high degree nodes are robust to reduced sampling, given that the duplicate sample count of a strain may be a proxy for its population prevalence. This is akin to the robustness of scale-free networks towards random failures, with sample count taking place of degree [15]. Consequently, increased surveillance has likely identified more strains of low prevalence given that low-degree strains are more likely to be removed from the network with random sub-sampling.

The network contains numerous cycles amidst its tree-like structure. Its 500 triangles indicate mutations at the same site between 3 sequences, while sparse squares indicate potential convergent evolution [11]. These structures are clearly displayed in genotype networks, while phylogenetic tree construction do not include convergent evolution structurally. The treelike topology of the network prevents longer cycles from forming. Further network summary statistics are shown in Table 2.1 for the enture network G and the giant component GC. The triangles are captured by global clustering C_{global}, which is equivalent to the proportion of triplets (3 connected nodes) that form a closed triangle.

	n	m	k_{mean}	k_{max}	diameter	C_{global}	ρ_k
$\boldsymbol{G}$	9714	7599	1.86	257	-	0.0096	-0.13
$\boldsymbol{GC}$	1629	2225	2.73	257	17	0.0010	-0.20

Table 2.1: Network statistics for entire network G and giant component GC.

The degree assortativity ρ_k represents the correlation between the degree of a node and its neighbors (Table 2.1). A negative value for both the entire network G and the giant component GC indicate that high degree nodes tend to attach to

low degree nodes, as is true conversely. Future investigation of vaccination strategies may consider potential high-degree hubs as immunization targets, given their high neighbor count and the transcendence of immunity in influenza A (H3N2) viruses, as well as potential bridges between highly connected regions of the network.

2.3.2 Network topology in time

The genotype network grows in time as new strains emerge and are sampled. The growth of the second largest component is shown in Figure 2.2, with each node colored by the first sample date of its corresponding strain. This component is large enough to span several years while remaining small enough to qualitatively observe network growth in time. The blue-shifted nodes represent the earliest observed strains among those belonging to this component, the first of which was sampled in late 2010. The majority of unique strains were sampled from 2012 to 2015, including multiple high-degree strains and their neighbors. The most recent strains from this network component are red-shifted, clearly depicting the tree-like growth process.

Numerous hubs are seen throughout the network, with the largest hubs existing around the 2012-2013 flu season that contributed numerous strains to this component (Figure 2.2, bottom). Seasonality is reflected in the sample date distribution of this component, with multiple peaks around the start of the calendar year during flu season.

Features of the genotype network remain fairly stable in time, even in the presence of a constantly increasing sampling rate. Genotype networks were constructed using samples within a 5 year window, sweeping across the entire sample set from 1999 to 2019. These temporally restricted genotype networks display the structure of the

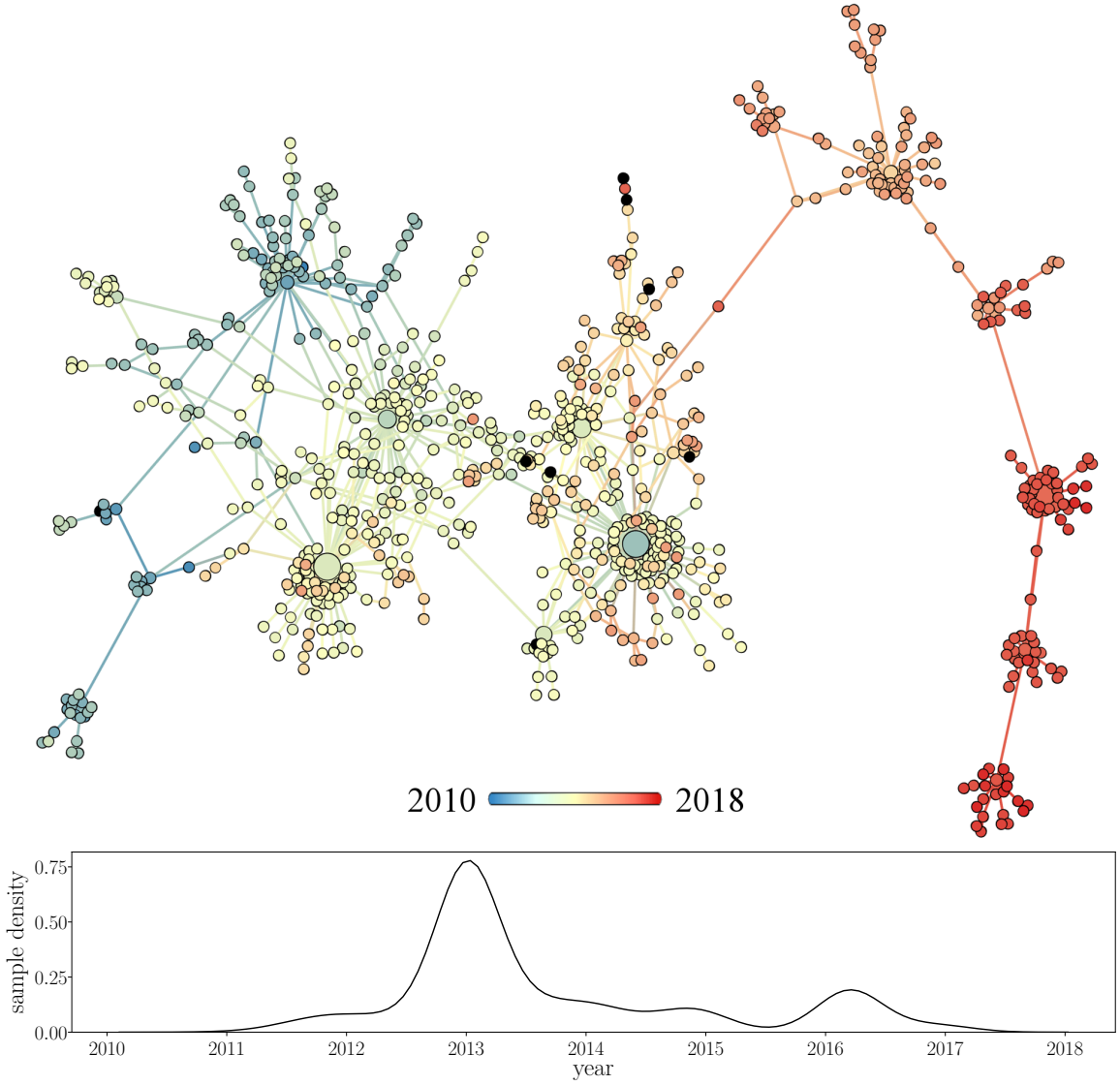


Figure 2.2: Sample dates among strains of second largest network component. **(top)** Nodes colored by first sample date (8 nodes with lacking sample dates colored black), with a larger radius corresponding to more samples (max sample count 337). **(bottom)** Sample date distribution across all dated samples of strains within the above network.

network local in time, an important consideration given that strains emerge and fall out of circulation. These networks display the increased availability of sequenced samples with each successive year, with notable increases in sampling since 2008

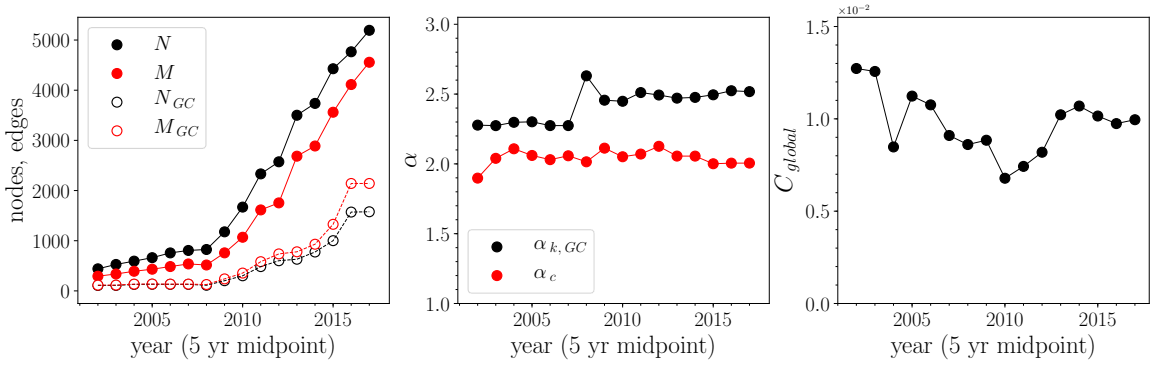


Figure 2.3: Network statistics in time. Influenza A (H3N2) HA genotype networks generated using samples within a sweeping 5 year window from July 1999 through June 2019, shown at midpoint. **(left)** Number of nodes and edges for entire network and giant component. **(center)** Power-law fit exponents α_c and $\alpha_{k,\,GC}$ with $c_{min} = k_{min} = 2$. **(right)** Global clustering coefficient C_{global} in time.

(Figure 2.3, left). The number of nodes and edges has grown steadily in the past 2 decades across both the entire network of the 5 year windows and its giant component.

Scaling of both degree distribution and component size distribution tails remain fairly constant in time. The power-law exponent for degree varied within $2.27 < \alpha_k < 2.63$, showing a modest increase in time, even as the network grew several times larger (Figure 2.3, center). Similarly, the power-law exponent for component size varied within $1.90 < \alpha_c < 2.13$, showing a weaker relationship with time and sample rate than α_k as the network grew (Figure 2.3, center). This shows that component size distribution approximately maintains the same scaling factor in time, while degree shows a slight increase in its scaling factor as more samples are contained within a 5 year window. This may result from the probable increase in the detection of low-degree nodes when sampling rates are high, shifting α_c to a steeper or higher value. Note that minimum values c_{min} and k_{min} were fixed at 2 to remove the effects of varying minimum values on the power-law fit on smaller networks, and may not be directly comparable to values shown in Figure 2.1 where $c_{min} = k_{min} = 7$.

Local cycles continue to remain prevalent in the network through time. The global clustering coefficient varied within $6.78\times10^{-3} < C_{global} < 1.27\times10^{-2}$, showing greater variability than scaling factors (Figure 2.2, right). Similarly, degree assortativity varied within $-0.365 < \rho_k < -0.124$, demonstrating variability but preserving the disassortative structure of the network. The above features demonstrate that in the presence of variable sequence sampling rates, genotype networks possess fairly consistent topological features that are highly predictable from recent years.

2.3.3 Epidemics in random graphs

To investigate how genotype network structure may influence the spread of disease, a multistrain SIRS model was constructed with an underlying network of strains (2.2.2: Methods). The incorporation of a genetic strain structure allows for both mutation between neighboring strains and cross-protective immune effects, existing as a function of genetic distance.

The connectivity or edge density of a genotype network may influence its endemic infection capacity, as suggested by cross-protective immune effects and the observed criticality within the genotype network structure. Here the effects of connectivity were investigated with the implementation of the multistrain model on $G(n, p)$ Erdős–Rényi (ER) random networks, for number of nodes n and edge probability p controlling connectivity. Endemic infection proportion I^* was observed across varying edge densities and levels of immunity transcendence to determine their effects on endemic infections for a genotype network of a given size.

Non-trivial dynamics are revealed by the multistrain epidemic model with an underlying genotype network structure of ER random networks. Endemic infec-

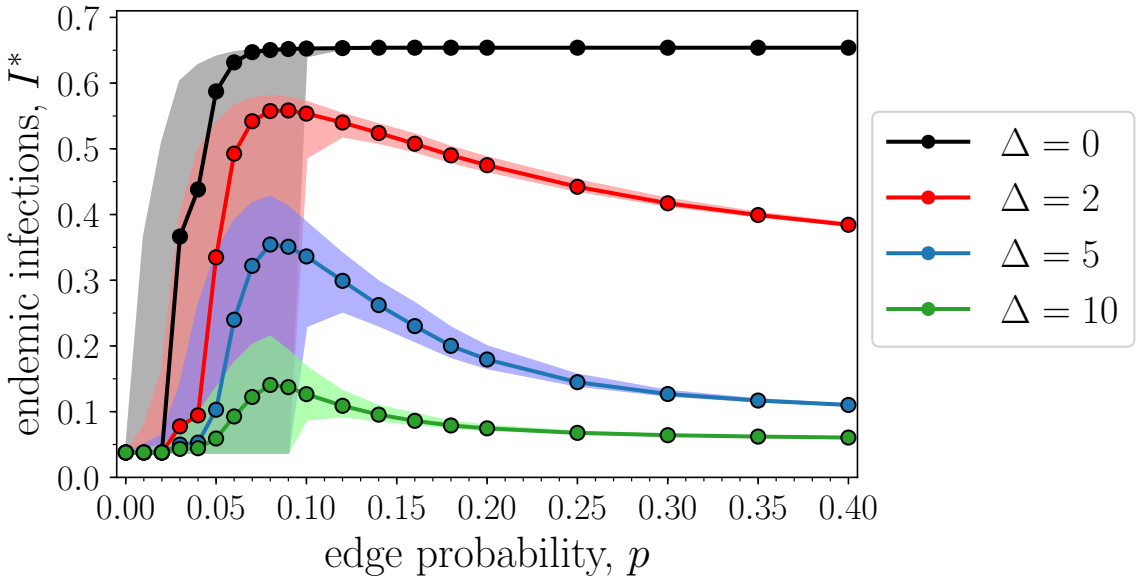


Figure 2.4: Endemic infections I^* as a function of ER random network edge density p and transcending immunity parameter Δ, using SIRS model with underlying genotype network. Median across 200 repetitions shown with 10th and 90th percentiles, shaded. Network size $n = 25$, mutation rate $\mu = 1/50$, transmission rate $\beta = 1/2$, recovery rate $\gamma = 1/6$, immune loss rate $\alpha = 1/100$.

tion proportions I^* are lowered in ER random genotype networks in the presence of high connectivity and non-zero transcending-immunity parameter Δ, producing cross-protective immune effects (Figure 2.4). Low connectivity resulted in lower I^* through increased fragmentation, resulting in numerous components that restrict mutation pathways between all strains. Together these dynamics produce an optimal connectivity that may be influenced by the parameterization of the multistrain model. An optimal edge density is observed at approximately $p = \frac{2}{n}$, close to that of the critical transition probability $p_c = \frac{1}{n-1}$ above which a giant component is expected. Around this critical transition, the random networks are known to have a power-law distribution of component sizes with exponent 1.5, close to the behaviour observed in Figure 2.3, center [16].

It is noted that the ER random networks are distinct in topology from the scale-free genotype networks observed for influenza A (H3N2) HA, and further exploration is required to understand the effects of edge density given different general network topologies.

2.4 Discussion

The influenza genotype networks explore a subspace of all networks that is predictable in structure as they grow in time. Features such as scale-free degree distributions and component size distributions remained present and fairly consistent in networks generated using temporal subsets of strain samples. This suggests that although future strains arise through stochastic mutation events, their effect on network structure may be predictable. Given the numerous mutations possible, it may not be possible to predict new strains with meaningful accuracy. However, it may be possible to predict their genetic relationship to strains existing in the network structure. Assuming the genetic distance is proportional to antigenic distance, this is a consequential development related to the understanding of cross-protective immune effects and vaccination strain selection.

The topology of influenza A (H3N2) HA genotype networks indicate that it may be influenced by strain-transcending immunity. This is further suggested by the dynamics of a multistrain epidemic model. The power-law degree fit is also steeper when lower degree nodes are consider ($k < 7$), indicating that high degree nodes are more prevalent than would be expected were there one scaling regime across all observed degrees. This corresponds to the presence of more hubs than would be

expected through degree-based preferential attachment. An possible explanation for this observation is that strain extinction prevents new nodes forming via mutation from older hubs, enabling the growth of new hubs.

The strong positive relationship between degree and sample count implies preferential attachment based on degree, however node age implements a consequential maximum age at which a node may be acquire new neighbors. This corresponds to the point at which the strain is not widely circulating or extinct in the host population. Furthermore, the multistrain model suggests that strain-transcending immunity drives this strain extinction process as cross-protective effects increase population immunity towards strains in time.

Together these observations suggest that influenza genotype networks explore a subspace predictable in structure, influenced by the effects of strain-transcending immunity. We call for the increased genomic surveillance of multistrain pathogens allow for similar analyses of other diseases with variable antigenic properties. Further analysis of influenza genotype networks may be considered for vaccine strain selection, analysis of evolutionary trajectory, and refinement of the understanding of cross-protective immunity.

Bibliography

[1] Wayan C W S Putri, David J Muscatello, Melissa S Stockwell, and Anthony T Newall. Economic burden of seasonal influenza in the United States. *Vaccine*, 36(27):3960–3966, Jun 2018.

[2] Noelle-Angelique M. Molinari, Ismael R. Ortega-Sanchez, Mark L. Messonnier, William W. Thompson, Pascale M. Wortley, Eric Weintraub, and Carolyn B. Bridges. The annual impact of seasonal influenza in the US: Measuring disease burden and costs. *Vaccine*, 25(27):5086 – 5096, 2007.

[3] Melissa A Rolfes, Ivo M Foppa, Shikha Garg, Brendan Flannery, Lynnette Bram-

mer, James A Singleton, Erin Burns, Daniel Jernigan, Sonja J Olsen, Joseph Bresee, and Carrie Reed. Annual estimates of the burden of seasonal influenza in the United States: A tool for strengthening influenza surveillance and preparedness. *Influenza and Other Respiratory Viruses*, 12(1):132–137, 01 2018.

[4] A Danielle Iuliano, Katherine M Roguski, Howard H Chang, David J Muscatello, Rakhee Palekar, Stefano Tempia, Cheryl Cohen, Jon Michael Gran, Dena Schanzer, Benjamin J Cowling, Peng Wu, Jan Kyncl, Li Wei Ang, Minah Park, Monika Redlberger-Fritz, Hongjie Yu, Laura Espenhain, Anand Krishnan, Gideon Emukule, Liselotte van Asten, Susana Pereira da Silva, Suchunya Aungkulanon, Udo Buchholz, Marc-Alain Widdowson, and Joseph S Bresee. Estimates of global seasonal influenza-associated respiratory mortality: a modelling study. *Lancet*, 391(10127):1285–1300, Mar 2018.

[5] Richard J. Whitley and Arnold S. Monto. Prevention and treatment of influenza in high-risk groups: Children, pregnant women, immunocompromised hosts, and nursing home residents. *The Journal of Infectious Diseases*, 194:S133–S138, 11 2006.

[6] Harish Nair, W Abdullah Brooks, Mark Katz, Anna Roca, James A Berkley, Shabir A Madhi, James Mark Simmerman, Aubree Gordon, Masatoki Sato, Stephen Howie, Anand Krishnan, Maurice Ope, Kim A Lindblade, Phyllis Carosone-Link, Marilla Lucero, Walter Ochieng, Laurie Kamimoto, Erica Dueger, Niranjan Bhat, Sirenda Vong, Evropi Theodoratou, Malinee Chittaganpitch, Osaretin Chimah, Angel Balmaseda, Philippe Buchy, Eva Harris, Valerie Evans, Masahiko Katayose, Bharti Gaur, Cristina O'Callaghan-Gordo, Doli Goswami, Wences Arvelo, Marietjie Venter, Thomas Briese, Rafal Tokarz, Marc-Alain Widdowson, Anthony W Mounts, Robert F Breiman, Daniel R Feikin, Keith P Klugman, Sonja J Olsen, Bradford D Gessner, Peter F Wright, Igor Rudan, Shobha Broor, Eric AF Simões, and Harry Campbell. Global burden of respiratory infections due to seasonal influenza in young children: a systematic review and meta-analysis. *The Lancet*, 378(9807):1917–1930, 2011.

[7] Yi Guan, Dhanasekaran Vijaykrishna, Justin Bahl, Huachen Zhu, Jia Wang, and Gavin J.D. Smith. The emergence of pandemic influenza viruses. *Protein and Cell*, 1(1):9–13, 2010.

[8] F. Carrat and A. Flahault. Influenza vaccine: The challenge of antigenic drift. *Vaccine*, 25(39-40):6852–6862, 2007.

[9] Scott E. Hensley. Challenges of selecting seasonal influenza vaccine strains for humans with diverse pre-exposure histories. *Current Opinion in Virology*, 8:85–89, 2014.

[10] Ben Peeters, Sylvia Reemers, Jos Dortmans, Erik de Vries, Mart de Jong, Saskia van de Zande, Peter J M Rottier, and Cornelis A M de Haan. Genetic versus antigenic differences among highly pathogenic H5N1 avian influenza A viruses:

consequences for vaccine strain selection. *Virology*, 503:83–93, Mar 2017.

[11] Andreas Wagner. A genotype network reveals homoplastic cycles of convergent evolution in influenza A (H3N2) haemagglutinin. *Proceedings of the Royal Society B: Biological Sciences*, 281(1786):20132763, 2014.

[12] Yun Zhang, Brian D Aevermann, Tavis K Anderson, David F Burke, Gwenaelle Dauphin, Zhiping Gu, Sherry He, Sanjeev Kumar, Christopher N Larsen, Alexandra J Lee, Xiaomei Li, Catherine Macken, Colin Mahaffey, Brett E Pickett, Brian Reardon, Thomas Smith, Lucy Stewart, Christian Suloway, Guangyu Sun, Lei Tong, Amy L Vincent, Bryan Walters, Sam Zaremba, Hongtao Zhao, Liwei Zhou, Christian Zmasek, Edward B Klem, and Richard H Scheuermann. Influenza research database: An integrated bioinformatics resource for influenza virus research. *Nucleic Acids Research*, 45(D1):D466–D474, 01 2017.

[13] Colin S. Gillespie. Fitting heavy tailed distributions: The poweRlaw package. *Journal of Statistical Software*, 64(2):1–16, 2015.

[14] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509, 10 1999.

[15] Duncan S. Callaway, M. E. J. Newman, Steven H. Strogatz, and Duncan J. Watts. Network robustness and fragility: Percolation on random graphs. *Physical Review Letters*, 85(25):5468–5471, 12 2000.

[16] M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 64:026118, Jul 2001.

Chapter 3

On the Approximate Bayesian Computation of Age-Weighted Preferential Attachment Models

Abstract

Real world networks may arise from undefined or partially understood generative processes. Proposed generative processes may be evaluated through the approximate Bayesian computation (ABC) of generative model parameters, identifying parameter values that lead to a desired network structure, if any exist. Here we investigate the use of ABC to identify parameter distributions that allow network generative models to reproduce some target network constructed by an unknown generative process. We ground this investigation in a real-world application by defining the target network to be a genotype network of the influenza A (H3N2) hemagglutinin (HA) surface

protein of antigenic and evolutionary significance. We show that (i) sufficient posterior distribution convergence is achieved across multiple generative models, (ii) nonlinear relationships exist between preference for degree and age when fitting a generative model to some target network, and (iii) age-weighted preferential attachment can improve the ability of a generative model to reproduce a viral genotype network. This shows that ABC is able to reverse-engineer the generative processes of a network through model evaluation, and that simple age-weighting functions may be used in combination with degree preference to create diverse network structures.

3.1 Introduction

3.1.1 Networks generative models

Network generative models are tools used to construct networks, explain features of existing networks, and give insight into the specific processes that govern the growth of networks. Networks are complex structures, with information contained in their topology that may be extracted via understanding the process by which the network formed. The degree distribution of a network alone may be used to broadly classify networks. A commonly used network generative model is the random graph, or Erdős-Rényi network [1]. In this generative model some number of nodes is specified, and all possible edges exists at some common probability, independent of one another. Erdős-Rényi networks have a characteristic binomial degree distribution that converges on Poisson in the limit of many nodes, producing a degree distribution that can be described in full by the mean of a Poisson distribution. A common generative model

with a notably different structure is the small-world network model [2, 3, 4, 5]. In small-world networks, a ring lattice is constructed to produce high local clustering, such that the neighbors of a node are likely to be connected to one another. Edges in the network are then rewired, replacing the end node of some edges with a random node. Depending on the proportion of edges rewired, the final network is often similar, except with a notable reduction in the average shortest path between nodes, hence the name small-world. Small-world networks models are capable of reproducing structure found in social networks and some neuronal networks of the brain [6, 7].

One of the more prominent categories of networks is those that take a heavy-tailed degree distribution, at times similar to a power law distribution among others [8, 9, 10]. Many real-world networks have a large number of nodes with low degree and a small number of nodes with high degree, such as the internet and electrical power grids [8]. This distribution may be found to remain scale-invariant, and in such a case is known as a scale-free network. The Barabási-Albert model offers an explanation for scale-free behavior with degree-based preferential attachment: a network is grown one node at a time, adding edges to existing nodes in proportion to their degree. The resultant network contains a scale-free degree distribution, in which a power law closely fits the distribution.

If a network is known to have been formed by some generative process, the corresponding model may be able to closely reproduce the network with some parameterization. Since generative processes are typically stochastic, this parameter may follow a distribution of values capable of producing some observed network. This parameter distribution may be inferred through Bayesian methods, given some observed network and the target. Bayesian inference is a power statistical method, gleaning information

on the conditional probability of an event as evidence becomes available.

3.1.2 Approximate Bayesian computation

In cases where a likelihood function is intractable, approximate Bayesian computation (ABC) may be used to approximate the distribution of a parameter [11, 12, 13, 14]. A versatile ABC algorithm is rejection sampling, in which parameter values are first drawn from some prior distribution to generate data. The generated data is then compared to the observed data with some summary statistic, and if the difference between summary statistics is below some level of error tolerance, the parameter value is added to the posterior distribution. This posterior distribution is an approximation, converging on the true posterior distribution with reductions in error tolerance.

The parameterization of a generative model may be determined through the use of rejection sampling ABC [15]. A network may be generated by a model under some parameters, and if the network is similar enough in structure to the target network, the parameters will be considered to belong to an approximation of the true posterior distribution. As the tolerance for similarity shrinks towards zero, the computed posterior converges on the true posterior for the parameters. This requires some summary statistic to compare the generated graph graph to the target. Numerous graph distances exist, such as graph edit distance, NetSimile, maximum common subgraph, and Laplacian spectral distance, among other spectral distances [16, 17, 18, 19]. Each distance captures the structural differences between graphs in different ways, with various algorithmic complexity: counting the number of nodes and edges that must be added or removed, performing a function on summary statistics such as degree and local clustering, or comparing the difference in eigenvalues of the adjacency or Lapla-

cian matrices of two networks [19]. A distance measure that is highly representative of the difference between between two graphs will produce greater confidence in the approximation of the posterior parameter distribution.

3.1.3 Genotype networks and age-weighted preferential attachment

Bayesian inference of network generative model parameters can be used to understand the mechanisms that produce a network, which is of particular importance when the mechanisms are largely unknown. ABC may be used to evaluate parameterizations leading to a particular network under a proposed generative model. The parameters and their approximated posterior values can inform the preference for attachment based on degree, as with nonlinear preferential attachment, or other parameters, such as preference based on the age of a node.

Here we conduct a practical application of ABC on network generative methods, for a genotype network of the hemagglutinin (HA) protein of influenza A (H3N2). In this network each node represents a unique sequence of HA, defined here as a strain, with edges connecting sequences that differ by one amino acid, indicating a plausible mutation pathway. The parameterizations of generative models that most closely fit this network offer insight on how it grows in time, and consequently, the evolution of this protein of the influenza A (H3N2) virus.

An important factor in the formation of genotype networks is time, as new strains are likely to be found in the genetic neighborhood of prevalent strains at a given time [20]. It may be possible to capture this behavior by using age-weighted preferential

attachment, by which the age of a node is a factor for how likely a new node is to attach to it [21]. A variety of age-weighted preferential attachment functions were explored to: (i) identify parameterizations of each model that best reproduce the target genotype network, (ii) identify relationships between parameters within models, and (iii) qualitatively compare the different generative methods, as well as understand a broader usage of ABC on network generative models [22].

3.2 Methods

3.2.1 Construction of the genotype network

We selected an influenza A (H3N2) hemagglutinin (HA) genotype network as the target network with which to explore the inference of generative model parameters, rooting the project in a real-world application. The H3N2 subtype has been a prominent pandemic strain of seasonal influenza since 2010, alongside H1N1 [23]. This network consists of strains of influenza A (H3N2), defined by unique protein sequences of the HA surface protein. Edges exist between strains whose sequences differ by just one amino acid, indicating a plausible mutation pathway. HA contains highly antigenic regions, or regions that the human immune system recognize and use as a target for antibody attachment. Thus the HA protein plays an important evolutionary role for influenza. Host immunity has been shown to drive the evolution of influenza away from past strains, indicating positive Darwinian selection in favor of antigenically novel strains [24]. This process, known as antigenic drift, necessitates annual updates to influenza vaccines that provide antigenic coverage of strains most

likely to be prevalent in the following year [25]. Similarly, it produces a growing network structure, dependent on the prevalent strains and host immunity within the population.

Influenza A (H3N2) HA genotype networks were demonstrated as having characteristic features, with a tree-like structure, heavy-tailed degree distributions, 3-cycles that indicate multiple mutations at the same amino acid site in the protein sequences, and 4-cycles that indicate convergent evolution, further observed in Chapter 2 [20]. Importantly, these networks are constantly changing in time, and only strains prevalent in the population are able to lead to new nodes via mutation. Although network generative models such as the Barabási-Albert model are able to reproduce preferential attachment, which may exist in genotype networks based on the viability or prevalence of a strain, preference based on the age of a node may better capture the dynamics underlying a viral genotype network [8]. Degree-based preferential attachment alone may be unable to reproduce the observed structure without consideration for node age. In particular, strains will not be sampled after some variable duration of existence. The unspecified generative process of genotype networks, their characteristic structure, and biological-informed suggestions for their generative processes make genotype networks an appropriate target network for Bayesian inference of network generative parameters.

Influenza A (H3N2) HA is chosen for genotype network construction due to not only its implication in human immune response, but also the availability of a large quantity of sequence data due to widespread and long-lasting surveillance efforts. Influenza sequences are constantly being added to publicly available databases, allowing for future analyses as more information becomes available. Here we extract sequence

data from the Influenza Research Database (IRD), a project funded by the National Institute of Allergy and Infectious Diseases [26].

From the IRD, influenza A (H3N2) HA sequences were obtained along with the followed information: the sequence of HA, sequence accession, collection date of each sample, country and state/province of origin, and strain name. Sequences were restricted to a length of 566 amino acids and of human origin (as influenza is a zoonotic disease and may be found in other organisms, however here we choose to ignore any evolution of HA that may have occurred in an animal reservoir). The data set consists of 30,175 samples obtained from January 1999 to August 2019. An incomplete year for 2019 was selected to account for any delays in data reporting at the time of acquisition in 2020. Across all samples, there were 2,762 missing amino acids (encoded as 'X') and 168 uncertain amino acids, which listed multiple amino acids possible as the true amino acid at a particular site (encoded as 'B','J', and 'Z'). Given that genotype network construction depends on a precision at the level of one amino acid, all sequences with missing or uncertain amino acids were removed from the data set ($n = 1,278$).

The resulting set of sequences consist of 28,880 samples that represent 9,714 unique sequences, indicating that the majority of samples consist of sequences already sampled (as strains are often sampled more than once). The samples are geographically concentrated on the United States, representing 66.6% of all samples. The country with the next highest sample count is Australia, at 5.6% of all samples, while the remaining 27.9% of samples are distributed among 79 countries. This geographic distribution does not accurately represent the distribution of seasonal influenza, instead biasing the data set towards the United States. This may affect the genotype net-

work by excluding prevalent strains in other regions. As such, the genotype network structure should be interpreted as having been generated primarily through strains observed in the dominant geographic regions of sampling.

3.2.2 Structure of the genotype network

The sequenced samples were then used to construct the genotype network. The nodes of a genotype network represent unique sequences, with edges existing between sequences that differ by exactly one amino acid. All pairs of sequences were compared to evaluate the presence of an edge between them, representing a plausible mutation pathway. Of the 9,714 unique sequences, 2,878 were isolated nodes with no edges between them and any other sequences. The remaining 6,836 sequences were connected to at least one other sequences, with 7,599 edges among them. These sequences formed 384 connected components, ranging in size from 2 to 1,629 nodes. The component size distribution is shown in Figure 3.1 (left) with a power-law fit. The power law was fit as per Alstott et al. 2014 [27], with a minimum component size $c = 7$ and a slope of $\alpha = 1.66$. Note that $P(C \geq c)$ spans approximately 2 orders of magnitude, with confidence in this scaling behavior dependent upon larger components that may result from more thorough strain surveillance coverage in the future.

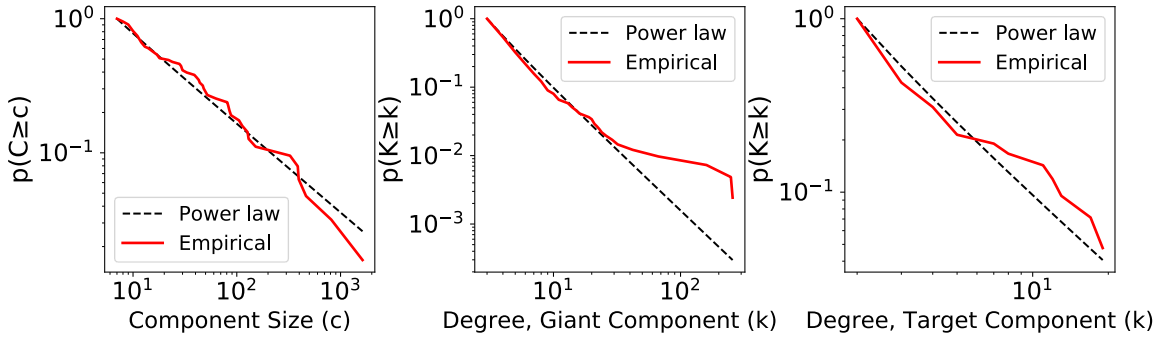


Figure 3.1: Complementary cumulative distribution of component size and degree. **(Left)** Component size distribution with power-law fit. Component size distribution approximately follows a power law for minimum component size $c_{min} = 7$ with slope $\alpha_c = 1.66$. **(Center)** Giant component (GC) degree distribution with power-law fit. GC degree distribution appears to have multiple scaling regimes, deviating from a power law for high degree nodes. Reference power law fit with $k_{min} = 3$ and slope $\alpha_k = 2.76$. **(Right)** Degree distribution of target component for ABC (8th largest component) resembles that of the GC, with a greater proportion of high degree nodes than the power-law fit. Reference power law fit with $k_{min} = 2$ and slope $\alpha_k = 2.29$.

The degree distribution of the giant component ($n = 1,629$) is heavy-tailed with multiple scaling regimes, as indicated by the poor fit of a power law across all degrees $k > 2$ (Figure 3.1, center). It does however fit well for all $k < 12$, with $\alpha_k = 2.76$. This suggests that preferential attachment by degree is likely not strictly responsible for how new nodes attach to existing nodes in this network, given the deviance from the expected fit for such a process when only one scaling regime is considered.

The target network component for ABC on the genotype network is restricted by computational limits, from the giant component of $n = 1,629$ to the largest component by which ABC across several generative models for numerous simulations was feasible. This results in a target network of size $n = 130$, the 8th largest component in the genotype network (Figure 3.2). This network has a degree distribution resembling that of the giant component, heavy-tailed with more high degree nodes than would be expected to be considered scale-free (Figure 3.1, right). Heavy-tailed

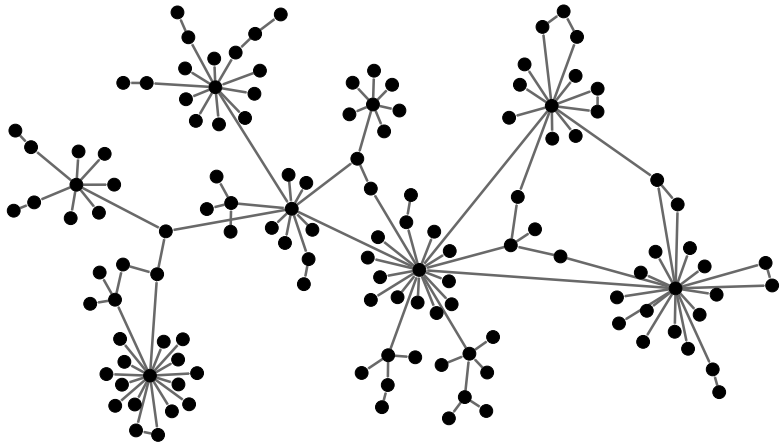


Figure 3.2: Target network used for ABC of generative model parameters, the 8th largest component of the influenza A (H3N2) HA genotype network ($n = 130$, $m = 139$).

distributions were found in components of all sizes, suggesting that the component used for ABC of generative models may be arbitrary, other than the possibility of increased variability in general structure as components become small.

The entire genotype network contains sparse cycles, with a global clustering coefficient $C_{global} = 0.0096$ indicating the rarity of 3-cycles. Similar levels of clustering are seen in the GC and the target network (Table 3.1). Nodes in the genotype network also tend to connect to nodes of dissimilar degree, referred to as disassortativity. This can be measured by the degree assortativity coefficient ρ_k ranging from -1 to 1, where negative values indicate disassortativity. Refer to Table 3.1 for further network statistics for the entire network, the GC, and the target component. The similarities of features between the target network and the GC suggest that parameterizations may be similar if the ABC is scaled up beyond computational limits. The target network preserves much of the structural features of the giant component: a tree like structure with a heavy-tailed degree distribution, numerous hubs, and a small number

	n	m	k_{mean}	k_{max}	diameter	clustering (C_{global})	assortativity (ρ_k)
G	9714	7599	1.86	257	-	0.0096	-0.13
GC	1629	2225	2.73	257	17	0.0010	-0.20
target	130	139	2.14	19	9	0.0015	-0.47

Table 3.1: Network statistics for entire network G, giant component GC, and the target network (8th largest component).

of cycles.

The components of a genotype network are determined largely by the coverage of sampling: in the presence of low sampling, fewer prevalent strains will be observed and the network will become more fragmented. Sequences within a component of this network are often separated in time on the order of weeks or months, while the average temporal differences between strains of different components was on the order of years. As such, the processes that lead to the emergence of genotype network structure are expected to remain more constant on a local time scale within a component than throughout the entire network. Annual sampling rates alone were several times higher in the latter portion of the data set from 2010-2019 than from 1999-2009, as influenza surveillance has increased in recent years. By selecting one component restricted in time, such a feature may be less impactful on the generative process. This justifies the decision to explore generative models that produce just one component.

3.2.3 Age-weighted network generative models

A number of network generative models are presented here, with justification for use with ABC based on the structure and biological context of the target network. A common restriction is made for the models with implications for edge density, in which the generative process consists of connecting a new node at each time step to

an existing node. This leads to a tree structure, but also forbids cycles. As such, the generative models cannot exactly replicate the cycles of the target network. This also enforces $n = m + 1$ to be true for the generated networks, thus $n \approx m$. However, these approximations may have minimal consequence, given the rarity of cycles in the target network and that $(n_{target} = 130) \approx (m_{target} = 139)$. The addition of extra edges at each time step according to some probability, either randomly or as some function of network structure, may be used in work beyond the scope of this paper to create a more versatile generative model.

The first generative model considered is the Barabási-Albert model [8]. This model constructs a network according to the principle of preferential attachment, implemented based on the degree of the nodes in a network at some time. The network is constructed by beginning with one node and adding one node at each time step t for $n_{final} - 1$ repetition. The probability of attaching to node i at t, $p_{i,t}$, is as follows:

$$p_{i,t} = \frac{k_{i,t}^{\alpha}}{\sum_j k_{j,t}^{\alpha}} \tag{3.1}$$

where $\alpha = 1$ for linear preferential attachment. In the ABC implementation of the Barabási-Albert model here, we define α as the parameter of interest, allowing it to vary. This allows for super-linear preferential attachment, where $\alpha > 1$, and sub-linear preferential attachment, where $\alpha < 1$. Although super-linear preferential attachment results in the dominance of attachment preference for one node as $n \to \infty$, the finite target network size justifies the consideration of this regime.

The Barabási-Albert model with non-linear preferential attachment model produces tree networks with heavy-tailed degree distributions, justifying its evaluation

for genotype network generation. It also serves as a framework for the models to follow. As discussed in Section 3.1.3, the age of a node in a genotype network has significance regarding the emergence of new nodes. As such, the subsequent generative models are preferential attachment models with the following form:

$$p_{i,t} = \frac{k_{i,t}^{\alpha}}{\sum_j k_{j,t}^{\alpha}} * \frac{f(\tau_{i,t})}{\sum_j f(\tau_{j,t})} \tag{3.2}$$

where $\tau_{i,t}$ is the age of node i at time t, and $f(\tau_{i,t})$ is some function of the age of the node that defines the preference for its attachment. If $f(\tau_{i,t})$ is some constant for all i, t, then the model is equivalent to the Barabási-Albert model with non-linear preferential attachment. However, we can define $f(\tau_{i,t})$ in a number of ways that may provide a framework to capture the generative process of genotype networks.

Equation 3.2 may thus be referred to as the age-weighted preferential attachment model, with the understanding that degree is an intrinsic factor beyond node age. We define the weight for age-weighted preferential attachment $\omega = f(\tau_{i,t})$, wherein the case $\omega = c$ for some constant c indicates no consideration for node age in attachment preference. The remaining generative models are variations of the age-weighted preferential attachment model in which define $f(\tau_{i,t})$ as a series of different functions, shown in Figure 3.3.

The first age-weighted generative model is threshold-based, defined by a step function $f_T(\tau)$:

$$f_T(\tau) = \begin{cases} 1 & \tau < T \\ 0 & \tau \geq T \end{cases}, \quad T \in [2, n_{final}] \tag{3.3}$$

such that nodes will be preferentially attached to with full age weight $\omega = 1$ until they reach age T, at which point they will no longer be attached to (Figure 3.3, left).

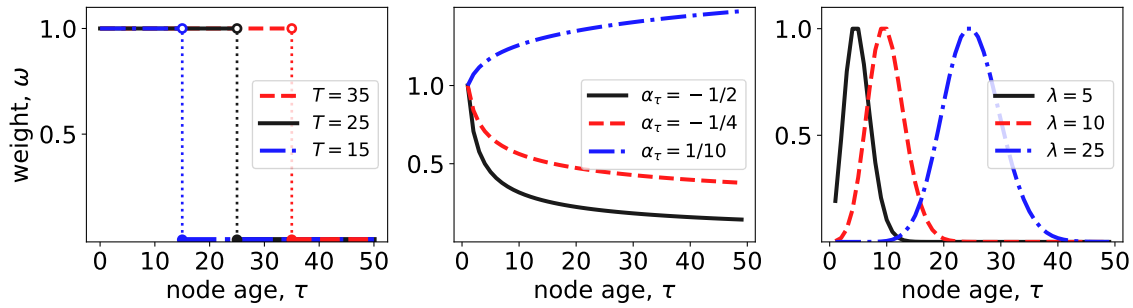


Figure 3.3: Age-based preferential attachment weight ω as a function of node age τ. **(Left)** Threshold-based weighting function f_T. Full preference ($\omega = 1$) for nodes younger than age threshold T, with no attachment for $\tau \geq T$ ($\omega = 0$). **(Center)** Power-law decay weighting: $\omega = \tau^{-\alpha}$. **(Right)** Poisson weighting, normalized by the maximum value of the Poisson PMF.

Only the most recent $T - 1$ nodes are considered for attachment. By varying T, new nodes will eventually ignore the high-degree founder nodes (when $\alpha > 0$) as they age out of the network's active region of preference. As $T \to n$, f_T will have no effect, but as $T \to 1$, the network will increase in diameter, becoming more chain-like.

The second age-weighted generative model is based on a power-law function $f_{PL}(\tau)$:

$$f_{PL}(\tau) = \tau^{\alpha_\tau} \ , \quad \alpha_\tau \in \mathbb{R} \tag{3.4}$$

by which the preference for a node is determined by its age raised to the exponent α_τ. Attachment based on site aging according to a power law was previously explored by Dorogovtsev & Mendes, 2000 [21]. Where $\alpha_\tau < 0$, the preference for a node decreases with its age (Figure 3.3, center). This inevitable reduction in attachment preference is the anticipated long-term behavior in the context of a genotype network. However, α_τ is not restricted to be negative, allowing for an increased attachment preference with age when $\alpha_\tau > 0$. Temporally local behavior in a genotype network could be

modeled by this, where as an emerging strain becomes more prevalent it becomes more likely to mutate into nearby strains.

The final age-weighting generative model mimics an epidemic curve, with function $f_P(\tau)$ based on the PMF of a Poisson distribution:

$$f_P(\tau) = \frac{\lambda^\tau e^{-\lambda}}{\tau!}\ , \quad \lambda > 0 \tag{3.5}$$

where λ is the mean of the Poisson distribution, and its floor $\lfloor\lambda\rfloor$ is the age at which preference is greatest (Figure 3.3, right). Thus f_P is a nonmonotonic function, allowing preference to increase before decreasing with age, the only such $f(\tau)$ considered here. The PMF of a Poisson distribution was selected as the shape of this curve to control its mean value with just one parameter, λ, simplifying the ABC process. Alternative nonmonotonic distributions may be considered beyond the scope of this paper.

Collectively, the above age-weighting functions (f_T, f_{PL}, f_P) account for different generative processes based on the age of a node, which in combination with degree-based preferential attachment form the generative models considered for comparison with the genotype network. These functions were selected to capture different mechanisms that may influence the structure of genotype networks, with tunable parameters that may be fit to a target network through methods of inference.

3.2.4 Approximate Bayesian computation for graphs

Posterior parameter distributions for the generative methods were approximated through a rejection algorithm ABC (Algorithm 1) [13]. ABC produces a set of sam-

ples with a distribution that approximately follows the true posterior, avoiding the need for an explicit likelihood function given its intractability in this application. A statistic must be defined that compares generated network H to target network G, such that the statistic is less than some error tolerance ϵ. Choice of this statistic must provide an accurate representation of the difference between G and H while remaining computationally efficient [28].

Algorithm 1: Rejection ABC for network generative models

Result: Approximated posterior parameter distribution(s)

$G \leftarrow$ target network;

$f(\alpha_k, \boldsymbol{\theta}) \leftarrow$ generative model;

$\epsilon \leftarrow$ error tolerance;

for $i \leftarrow 1$ **to** N **do**

  sample $a_k^* \sim \text{Prior}(a_k)$;

  for θ *in* $\boldsymbol{\theta}$ **do**

    sample $\theta^* \sim \text{Prior}(\theta)$;

  end

  $H \leftarrow f(a_k^*, \boldsymbol{\theta}^*)$;

  if *GraphDistance(G,H)* $< \epsilon$ **then**

    append a_k^* to Posterior(a_k);

    for θ *in* $\boldsymbol{\theta}$ **do**

      append θ^* to Posterior(θ);

    end

  end

end

In this application the ABC statistic is a graph distance measure, used to compare the similarity of two networks. Here graph distance is the distance between two networks, while network distance is the distance between two nodes within a network. Numerous graph distance measures exist, whose exact form must be chosen carefully for use with ABC. In Wills & Meyer 2020, graph distance metrics are evaluated on their ability to distinguish a null population of graphs from an alternative population[19]. They show that among a number of distance measures, a preferential attachment network is best distinguished from an Erdős-Rényi random graph by the combinatorial Laplacian spectral distance followed by the adjacency spectral distance. However, adjacency outperforms Laplacian spectral distance when a partial number of eigenvalues are considered. Although preferential attachment networks are poorly distinguishable from configuration models, in which degree distributions are equivalent but attachments are otherwise random, spectral distances outperform edit distance, DeltaCon, and NetSimile [29, 17, 19].

For these reasons, the graph distance used here with ABC is the truncated adjacency spectral distance, d_A:

$$d_A(G, H) = \sqrt{\sum_{i=1}^{k} (\lambda_{i,\, A_G} - \lambda_{i,\, A_H})^2}\ , \quad k < n \tag{3.6}$$

for networks G, H with n nodes, where A_G, A_H are the adjacency matrices and λ_i is the i^{th} eigenvector of an adjacency matrix. Here only the first k eigenvalues are compared between graphs, with a fixed $k = 10$ for all analyses that follow. Setting $k < n$ allows the distance metric to ignore the fine local structure, which may be unnecessary for comparison here given the stochastic nature of the generative processes [19]. The error tolerance ϵ of d_A is explored at multiple values for a given set of

generated networks. Values of ϵ are chosen to demonstrate a graphical convergence of posterior distribution, such that the smallest ϵ shown in the results provides the best approximation of the posterior.

3.3 Results

3.3.1 Degree-only preferential attachment

ABC was used to compute the posterior distribution for α_k of the non-linear preferential attachment model for the target genotype network component (Figure 3.4). A prior of $\alpha_k \sim U(0, 4)$ was defined. Convergence of the posterior distribution of α_k is demonstrated in Figure 3.4, center, as the error tolerance is reduced to $\epsilon = 1.50$. This strict error tolerance accepted 1.48% of the values drawn from the prior for α_k across 5000 repetitions. Under $\epsilon = 1.50$ we observe an unimodal distribution for α_k with mean 1.06 and standard deviation 0.149. As α_k is reduced from its mean we see a small penalty in the sublinear regime, while increasing α_k into the highly superlinear region results in notably different networks than the target (Figure 3.4, right).

The preferential attachment model based strictly on degree shows a slightly non-linear preference for attachment, with $\bar{\alpha}_k = 1.06$. The mean of this distribution increased with a reduction in ϵ, and may be slightly higher in the limit of smaller error tolerances. The prior for α_k was sufficient in accommodating values that provided the closest fit between the generated network and the target under this model.

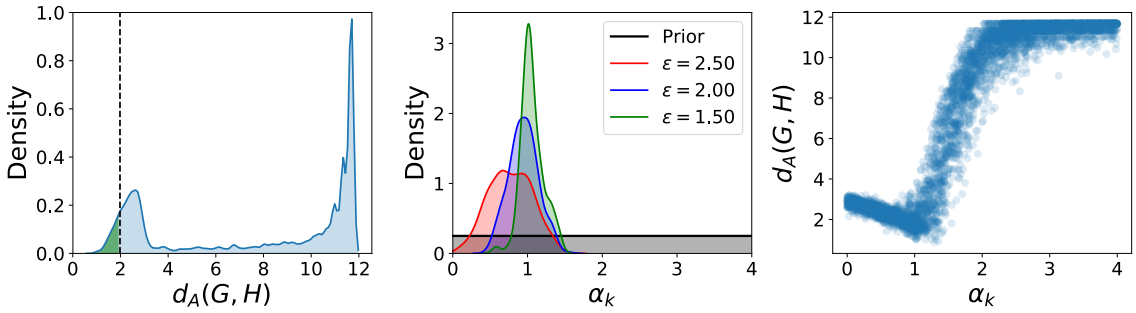


Figure 3.4: Non-linear preferential attachment model for target network G. Fixed $n = 130$. **(Left)** Distribution of truncated adjacency spectral graph distance $d_A(G, H)$ between G and generated network H. Error tolerance $\epsilon = 1.50$ shown, with 1.48% (74/5000) of prior parameter samples below ϵ. **(Center)** Posterior distribution of α_k demonstrates convergence as $\epsilon \to 0$, with prior $\alpha_k \sim U(0, 4)$. Mean of most restrictive posterior $\bar{\alpha}_{k,\epsilon=0.05} = 1.06$ ($\sigma = 0.149$). **(Right)** $d_A(G, H)$ as a function of α_k is reduced near $\bar{\alpha}_{k,\epsilon=0.05}$, with a greater distance penalty in the most extreme super-linear regime than the sub-linear regime.

3.3.2 Threshold age-weighting

Age-weighting was then introduced with the maximum age threshold model. Priors were defined as $\alpha_k \sim U(0, 4)$ and maximum age $T \sim U(2, N)$. The error distribution across all repetitions is shown in Figure 3.5, left. Convergence of the posterior distribution of α_k is demonstrated in Figure 3.5, center-left, as the error tolerance is reduced to $\epsilon = 0.75$. This strict error tolerance accepted 1.24% of the values drawn from the prior for α_k and T across 10^4 repetitions. Under $\epsilon = 0.75$ we observe an wide distribution for α_k with mean 5.82 and standard deviation 2.20. Unlike α_k, we see T converge on a narrow distribution with mean 18.66 and standard deviation 2.41 (Figure 3.5, center-right).

The lack of convergence on a tight distribution for α_k is explained by the relationship between α_k and T, shown in Figure 3.5, right. Generated networks could be produced with a moderate fit for $1 < \epsilon < 2$ for a wide range of T, however the

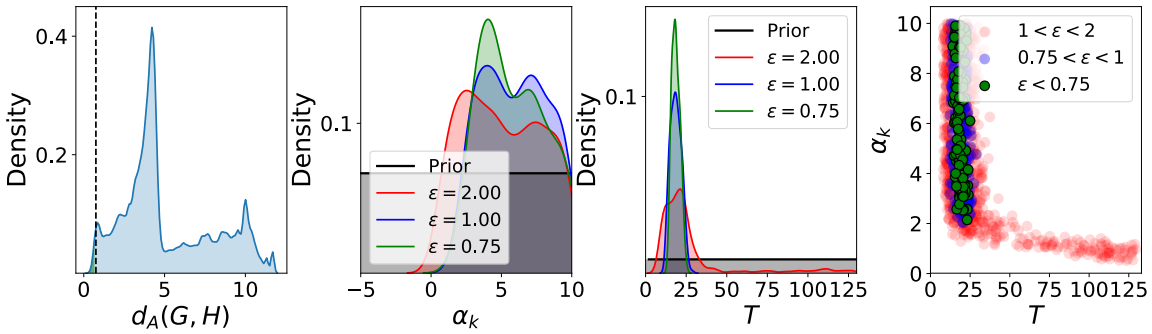


Figure 3.5: Preferential attachment model with threshold-based age-weighting according to f_T for target network G. Fixed $n = 130$. **(Left)** Distribution of $d_A(G, H)$ between G and generated network H. Error tolerance $\epsilon = 0.75$ shown. **(Center-Left)** Posterior distributions of exponent α_k demonstrate convergence towards a wide distribution of values as $\epsilon \to 0$. Mean of most restrictive posterior $\bar{\alpha}_{k,\epsilon=0.75} = 5.82$. **(Center-Right)** Posterior distributions of maximum age threshold T demonstrate convergence towards a narrow range of values centered around $\bar{T} = 18.66$ ($\sigma = 2.41$). **(Right)** Accepted values of α_k and T at varying ϵ levels indicate a strong relationship for acceptance to the posterior. A wide range of values may be accepted to the posterior for α and T under a large error tolerance, dependent upon the value of the other parameter.

best fitting networks were dependent upon T close to its restrictive posterior mean of 18.66. For this value of T, α_k could take a wide range of values, all in the super-linear regime of preferential attachment by degree. The effect this has on the network structure is the formation of numerous hubs with $k < T$, preventing the super-linear parameterization from condensing the network into one large hub. Note that when $T = n = 130$ this model is equivalent to the non-linear preferential attachment model without age-weighting. The generated networks most similar to the target were produced with T notably dissimilar to n, suggesting that the inclusion of a maximum age for node attachment may better capture the structure of the genotype network.

3.3.3 Power law age-weighting

Age-weighting was then explored with a power-law function of age. Priors were defined as $\alpha_k \sim U(-10, 15)$ and $\alpha_\tau \sim U(-15, 15)$. The error distribution across all repetitions is shown in Figure 3.6, left. There is considerable convergence of the posterior distribution of α_k on zero as the error tolerance is reduced to $\epsilon = 1$ (Figure 3.6, center-left). This strict error tolerance accepted 0.50% of the values drawn from the prior for α_k and α_τ across 10^4 repetitions. Unlike α_k, we see α_τ converge on high positive values (Figure 3.6, center-right).

The power law function of age produces an interesting effect where well-fitting models nearly eliminate preferential attachment based on degree. A value of $\alpha_k = 0$, which is converged upon, demonstrates no weighting effects from the degree of a node. Instead, we see a highly non-linear preference for older nodes, given the high

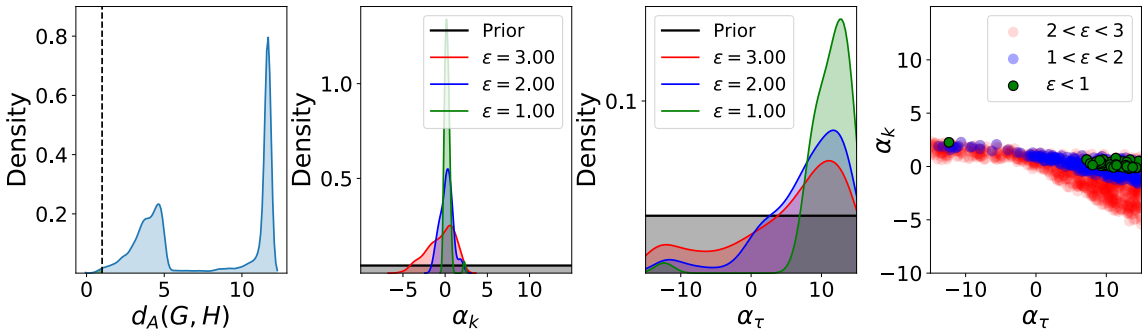


Figure 3.6: Preferential attachment model with power-law age-weighting according to f_{PL} for target network G. Fixed $n = 130$. **(Left)** Distribution of $d_A(G, H)$ between G and generated network H. Error tolerance $\epsilon = 1$ shown. **(Center-Left)** Posterior distributions of exponent α_k demonstrate convergence towards 0 as $\epsilon \to 0$. Mean of most restrictive posterior $\bar{\alpha}_{k,\epsilon=1} = 0.25$. **(Center-Right)** Posterior distributions of age exponent α_τ demonstrate convergence towards high values with $\bar{T} = 11.15$ ($\sigma = 3.95$). **(Right)** Accepted values of α_k and α_τ at varying ϵ levels indicate a relationship for acceptance to the posterior. A wide range of values may be accepted to the posterior for α_τ under a narrow range of acceptable α_k.

and positive α_τ posterior values. Interestingly, a small number of highly negative α_τ exist in the posterior, existing alongside non-zero α_k. Positive values of α_k are associated with α_τ in the posterior, while zero or negative values of α_k are associated with a better fit when α_τ is highly positive.

3.3.4 Poisson age-weighting

The final age-weighting scheme is the Poisson function of age. Priors were defined as $\alpha_k \sim U(-5, 10)$ and $\lambda \sim U(2, N)$. The error distribution across all repetitions is shown in Figure 3.7, left. Convergence of the posterior distribution of α_k is demonstrated in Figure 3.7, center-left, as the error tolerance is reduced to $\epsilon = 0.75$. This strict error tolerance accepted 0.46% of the values drawn from the prior for α_k and λ across 5000 repetitions. Under $\epsilon = 0.75$ we observe an wide distribution for α_k with mean 2.60 and standard deviation 2.23, skewed towards positive values. We see T fail to converge on a narrow distribution, with mean 77.48 and $\sigma = 41.9$ (Figure 3.7, center-right).

Although the posterior distributions do not converge on narrow distributions, they do converge on a narrow set within their 2-dimensional space, shown in Figure 3.7, right. The posterior values of α_k and λ are closely related, where high values of α_k are associated with small values of λ, and low values of α_k are associated with high values of λ. A wide range of λ are associated with a narrow range of α_k from about 0 to 3. This suggests that the delay and duration at which a node is preferentially attached to interacts considerably with the preference for attachment based on degree for a given network structure. As with the threshold-based age function, we see that considerable super-linear degree preference is allowable when preference for age is

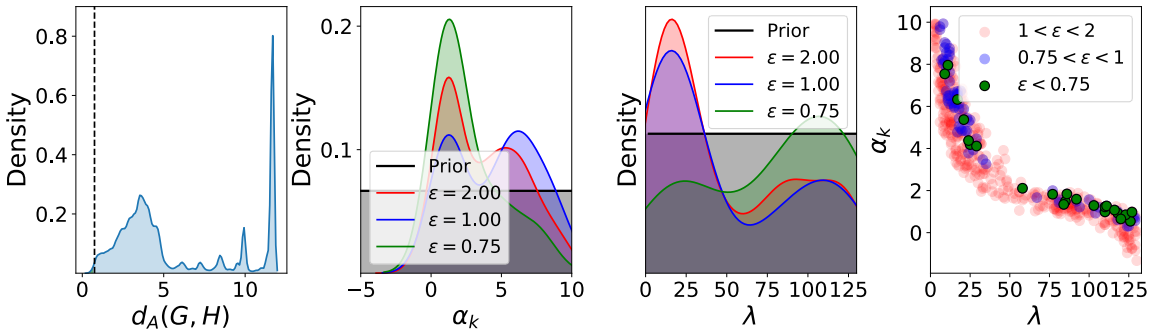


Figure 3.7: Preferential attachment model with Poisson age-weighting according to f_P for target network G. Fixed $n = 130$. **(Left)** Distribution of $d_A(G, H)$ between G and generated network H. Error tolerance $\epsilon = 0.75$ shown. **(Center-Left)** Posterior distributions of exponent α_k demonstrate convergence towards a wide distribution of values as $\epsilon \to 0$, skewed towards high positive values. Mean of most restrictive posterior $\bar{\alpha}_{k,\epsilon=0.75} = 2.60$. **(Center-Right)** Posterior distributions of λ demonstrate convergence towards a broad bimodal distribution. Moderately selective error tolerances preserve a peak around $\lambda = 20$, while the most restrictive posterior is comprised mostly of $\lambda > 75$. **(Right)** Accepted values of α_k and λ at varying ϵ levels indicate a strong non-linear relationship for acceptance to the posterior. A wide range of values may be accepted to the posterior for α and λ under a larger error tolerance, dependent upon the value of the other parameter.

restricted to values significantly less than the size of the entire network.

3.4 Discussion

The above presents an analysis of various age-weighting schemes as applied to the task of replicating a genotype network with an undefined generative process. The use of ABC provided insight on the parameterizations of these models that can closely replicate the genotype network, with said parameters and their values of potential use in understanding the evolution of the antigenically significant influenza A (H3N2) HA protein. Of significance was the use of ABC for network generative models, in combination with a truncated adjacency spectral distance, that demonstrated notable convergence of the posterior parameter distributions in the limit of a sufficient

number of prior draws. This suggests that ABC may be a useful method for reverse-engineering the generative processes of networks, by identifying parameters that can reproduce a target network under some generative model, informed by relevant or proposed processes pertaining to the target network.

The inclusion of age-weighted preferential attachment was shown to have a considerable effect on the generative process while fitting to the target genotype network. Each age-weighting function explored demonstrated a characteristic relationship between age and degree that generated similar network structures across different parameter values. Introducing a maximum age for attachment allowed highly super-linear preference based on degree, but only in the limit of a maximum age that was a small fraction of the total age of the final network. A power law function of age largely nullified the effects of degree-based preference, setting degree weights to be nearly constant in favor of super-linear age-based preferential attachment. Poisson age-weighting resulted in a model where the entire prior distribution of the Poisson mean λ could be accepted, but only for a narrow distribution of α_k dependent upon λ. This function was the most biologically informed age-weighting scheme, with its nonmonotonic function of age simulating an epidemic curve of strain prevalence.

These generative models demonstrate that a variety of parameterizations can produce the same network structure. With an understanding of the relationship between parameters, as shown, knowledge about the network and all but one parameter can be used to identify a distribution for an unknown generative parameter. This suggests that ABC is a useful method in the inference of network generative model parameters, that age-weighting may be a useful consideration for generative models, and that the processes which lead to a network's structure may be identified and validated with

the use of an informed generative model.

Bibliography

[1] E. N. Gilbert. Random graphs. *Ann. Math. Statist.*, 30(4):1141–1144, 1959.

[2] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of small-world networks. *Nature*, 393(6684):440–442, 1998.

[3] M.E.J. Newman and D.J. Watts. Renormalization group analysis of the small-world network model. *Physics Letters A*, 263(4):341 – 346, 1999.

[4] M. E. J. Newman and D. J. Watts. Scaling and percolation in the small-world network model. *Physical Review E*, 60:7332–7342, Dec 1999.

[5] A. Barrat and M. Weigt. On the properties of small-world network models. *The European Physical Journal B - Condensed Matter and Complex Systems*, 13(3):547–560, 2000.

[6] L. A. N. Amaral, A. Scala, M. Barthélémy, and H. E. Stanley. Classes of small-world networks. *Proceedings of the National Academy of Sciences*, 97(21):11149–11152, 2000.

[7] Danielle Smith Bassett and Ed Bullmore. Small-world brain networks. *The Neuroscientist*, 12(6):512–523, 2006.

[8] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509, 10 1999.

[9] Albert-László Barabási and Eric Bonabeau. Scale-free networks. *Scientific American*, 288(5):60–69, 2003.

[10] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Physical Review Letters*, 86:3200–3203, Apr 2001.

[11] Mark A. Beaumont. Approximate Bayesian computation in evolution and ecology. *Annual Review of Ecology, Evolution, and Systematics*, 41:379–406, 2010.

[12] Katalin Csilléry, Michael G.B. Blum, Oscar E. Gaggiotti, and Olivier François. Approximate Bayesian computation (ABC) in practice. *Trends in Ecology & Evolution*, 25(7):410 – 418, 2010.

[13] Jarno Lintusaari, Michael U. Gutmann, Ritabrata Dutta, Samuel Kaski, and Jukka Corander. Fundamentals and Recent Developments in Approximate Bayesian Computation. *Systematic Biology*, 66(1):e66–e82, 09 2016.

[14] Jean-Michel Marin, Pierre Pudlo, Christian P. Robert, and Robin J. Ryder. Approximate Bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180, 2012.

[15] Oliver Ratmann, Ole Jørgensen, Trevor Hinkley, Michael Stumpf, Sylvia Richardson, and Carsten Wiuf. Using likelihood-free inference to compare evo-

lutionary dynamics of the protein networks of H. pylori and P. falciparum. *PLoS Computational Biology*, 3(11):e230–e230, 11 2007.

[16] Xinbo Gao, Bing Xiao, Dacheng Tao, and Xuelong Li. A survey of graph edit distance. *Pattern Analysis and Applications*, 13(1):113–129, 2010.

[17] Tina Eliassi-Rad Christos Faloutsos Michele Berlingerio, Danai Koutra. Netsimile: A scalable approach to size-independentnetwork similarity. *ArXiv:1209.2684*, 2012.

[18] Horst Bunke and Kim Shearer. A graph distance metric based on the maximal common subgraph. *Pattern Recognition Letters*, 19(3):255 – 259, 1998.

[19] Peter Wills and François G. Meyer. Metrics for graph comparison: A practitioner's guide. *PLoS ONE*, 15(2):1–54, 02 2020.

[20] Andreas Wagner. A genotype network reveals homoplastic cycles of convergent evolution in influenza A (H3N2) haemagglutinin. *Proceedings of the Royal Society B: Biological Sciences*, 281(1786):20132763, 2014.

[21] S. N. Dorogovtsev and J. F. F. Mendes. Evolution of networks with aging of sites. *Physics Review E*, 62:1842–1845, Aug 2000.

[22] Christian P. Robert, Jean-Marie Cornuet, Jean-Michel Marin, and Natesh S. Pillai. Lack of confidence in approximate bayesian computation model choice. *Proceedings of the National Academy of Sciences*, 108(37):15112–15117, 2011.

[23] CDC. Estimates of the incidence of symptomatic influenza by season and age-group, United States, 2010–2016. In *"Key Facts About Influenza"*, 2019. https://www.cdc.gov/flu\/about/keyfacts.htm.

[24] W M Fitch, J M Leiter, X Q Li, and P Palese. Positive Darwinian evolution in human influenza A viruses. *Proceedings of the National Academy of Sciences*, 88(10):4270–4274, 1991.

[25] G W Both, M J Sleigh, N J Cox, and A P Kendal. Antigenic drift in influenza virus H3 hemagglutinin from 1968 to 1980: multiple evolutionary pathways and sequential amino acid changes at key antigenic sites. *Journal of Virology*, 48(1):52–60, 1983.

[26] Yun Zhang, Brian D Aevermann, Tavis K Anderson, David F Burke, Gwenaelle Dauphin, Zhiping Gu, Sherry He, Sanjeev Kumar, Christopher N Larsen, Alexandra J Lee, Xiaomei Li, Catherine Macken, Colin Mahaffey, Brett E Pickett, Brian Reardon, Thomas Smith, Lucy Stewart, Christian Suloway, Guangyu Sun, Lei Tong, Amy L Vincent, Bryan Walters, Sam Zaremba, Hongtao Zhao, Liwei Zhou, Christian Zmasek, Edward B Klem, and Richard H Scheuermann. Influenza research database: An integrated bioinformatics resource for influenza virus research. *Nucleic Acids Research*, 45(D1):D466–D474, 01 2017.

[27] Jeff Alstott, Ed Bullmore, and Dietmar Plenz. Powerlaw: A python package for analysis of heavy-tailed distributions. *PLoS ONE*, 9(1):e85777, Jan 2014.

[28] Voss J. Webster M. Barber, S. The rate of convergence for approximate bayesian

computation. *Electron. J. Statist.*, 9(1):80–105, 2015.

[29] Danai Koutra, Joshua T. Vogelstein, and Christos Faloutsos. Deltacon: A principled massive-graph similarity function. *Proceedings of the 2013 SIAM International Conference on Data Mining*, pages 162–170, 2013.

Chapter 4

On Optimal Multivalent Vaccination Strategies on Viral Genotype Networks

Abstract

We implement a novel approach for vaccine strain selection based on a genotype network of viral strains. Current strategies for selecting vaccine strains of multistrain pathogens involve present and forecasted incidence of particular strains. Here we emphasize the effects of transcending immunity, and exploit the genetic similarity between strains to determine optimal strategies in the case of multistrain vaccination. We employ a genetic algorithm (GA) to find optimal strategies in the $\binom{N}{k}$ search space of k vaccines on N strains, seeking to reduce the number of strains that may be reached in an outbreak. We tested the strategy on toy networks of varying size

and structure, before searching optimal strategies for multiple real-world influenza A (H3N2) genotype networks. This approach consistently reduced the mean expected outbreak size by strain count, with significant improvements on random searches. Evolved solutions were evaluated on influenza A genotype networks that grew beyond the time of solution computation, simulating the 6 month delay between strain selection and distribution. Despite ignorance toward future states of the genotype network, GA-evolved strategies consistently outperformed even the best random solutions after a year of novel strain emergence. Our approach suggests that knowledge of the genotype network can provide useful insight for vaccine strain selection.

4.1 Introduction

High mutation rates in RNA viruses such as the Zaire ebolavirus [1], influenza A virus [2], and rabies lyssavirus [3] lead to numerous contemporaneous strains [**?**]. Vaccines are developed based on the antigenic properties of such viruses, however vaccine effectiveness can be less than ideal: influenza vaccine efficacy has been approximately 40% since 2005 [4, 5, 6, 7, 8]. Effective vaccination is challenged by: (i) rapid evolution of viruses away from the antigenic properties of strain(s) used for vaccines [9], and (ii) properly selecting strains for vaccines such that antibodies have a wide-reaching effect on prevalent and future strains [10, 11].

Here, we address the problem of selecting vaccination strains that provide maximal antigenic coverage, in the case where multiple vaccination strains may be used. Identifying which strains to include in a vaccine is a problem with time complexity $\mathcal{O}\binom{N}{k}$ for N strains in the population and k chosen vaccination strains. For large enough

N and even modest increments in k, the time to brute-force an optimal combination of vaccination strains could be infeasible, especially with the use of in-depth modeling with compartmental or agent-based models, let alone laboratory viral inhibition assays.

Each spring and fall, the World Heath Organization (WHO) makes recommendations for specific strains to be included in the influenza vaccine for each hemisphere. WHO bases their recommendations largely on the current and forecasted incidence of a particular strain in the upcoming flu season, as well as the availability of similar vaccine viruses [12]. Although attention is given to the genetic similarity between strains by incorporating phylogenetic analysis, the information contained within genotype networks and complementary network analyses might not fully exploited. While the WHO typically only recommends one or two vaccine strains per subtype of influenza, we explore a situation in which multiple vaccine strains ($k \geq 3$) are considered for a viral subtype using its genotype network, given the history of poor vaccine efficacy. Our approach suggests that choosing multiple strains based on knowledge of the network structure can greatly increase the efficacy of a vaccine.

We developed an approximation of vaccine efficacy through suppression of outbreak potential in the presence of vaccinated strains. Transcending effects of immunity, observed in viruses such as influenza [13], allow for genetically similar strains to be influenced by nearby vaccines. A genotype network was used to model the genetic similarity between strains, allowing for real-world and simulated network structure to be evaluated.

In this paper we implement a genetic algorithm to find ideal vaccination strains for a given genotype network. In Section 2 we discuss the details of the GA im-

plementation, including solution representation, fitness evaluation, and the use of genotype networks. In Section 3, we first test this approach on a series of simple toy networks and a small Erdős-Rényi random graph to provide a clear understanding of how the vaccination strategy evolves on relatively simple network structures. We then apply the GA to a series of influenza A H3N2 genotype networks of ranging in size and complexity from size 81 to 1430, to test the approach on complex and large real world genotype networks. Finally, we evaluated GA-evolved and random vaccination strategies on an influenza network that is growing through the addition of novel strains arising via mutation over time, to simulate the lag in time between the selection of the vaccination strain and the end of a flu season.

4.2 Methods

4.2.1 Genotype network

A given set of strains are related to one another through a genotype network. Each node in this network corresponds to a unique gene or protein sequence (defined as a strain), with edges existing between strains whose sequences differ by one base pair or amino acid (indicating a plausible mutation pathway). In this paper, sequences will be assumed to be the amino acids of a specified antigenic protein. In the real-world application, this will be the hemagglutinin (HA) surface protein of influenza A H3N2.

4.2.2 Outbreak fitness function

In simplistic epidemic models the basic reproductive number R_0 may be used to determine epidemic phase transitions, with $R_0 = \frac{\beta}{\lambda}$ where β is the number of new cases generated by a case in time step t and $\frac{1}{\gamma}$ is the mean time steps of infectivity for a case. In an infinite well-mixed homogeneous population, R_0 is the expected number of new infections each individual case will produce. For $R_0 < 1$, the number of cases of disease is expected to approach 0 in time, but for $R_0 > 1$ sustained transmission is expected; thus $R_0 = 1$ represents the epidemic threshold.

Here we define R_0^{eff} as the normalized effective R_0 after the effects of vaccination, such that for strain i and set of vaccination strains V:

$$R_0^{eff}(i) = \begin{cases} 1 & \text{if } V = \emptyset \\ 0 & \text{if } i \in V \\ \prod_{v \in V} (1 - e^{-x_{iv}/\Delta}) & \text{otherwise} \end{cases} \tag{4.1}$$

where x_{iv} is the genetic distance between strains i, v, and Δ is the tunable transcendence of immunity parameter. Genetic distance is determined from the shortest path in the network, which was observed to closely approximate genetic distance in real-world influenza networks. In the evaluation of fitness on a growing network component in Section 3, we allow the final network distances to be used in the calculation of an incomplete network. $R_0^{eff} = 1$ in the absence of any vaccines, but is reduced to 0 for directly vaccinated strains, and otherwise equals the product of immunity that transcends vaccinated strains as a log decaying function of genetic distance.

We define R_0^{crit} as the normalized epidemic threshold, constrained to $(0, 1)$ for all

$R_0 > 1$:

$$R_0^{crit} = \frac{1}{R_0} \text{ , for } R_0 > 1 \tag{4.2}$$

In this paper we let $R_0 = 2$, a value comparable to that of Ebola and pandemic influenza, such that $R_0^{crit} = \frac{1}{2}$.

The fitness F for a given set of vaccination strains V on network G is found by: (i) removing subcritical strains ($R_0^{eff}(i) < R_0^{crit}$), which potentially (and ideally) fragments the network into multiple components, then (ii) computing the mean component size for each strain i:

$$F(V, G) = \frac{\sum_j (j_n)^2}{G_n^2} \text{ for component size } j_n, \text{ network size } G_n \tag{4.3}$$

Thus $F(V, G)$ is the expected number of super-critical strains an outbreak can reach through known strains: it is the expected component size of an outbreak at a random strain. Minimizing this value will reduce the number of known strains an outbreak will reach, and necessitate evolutionary detours around vaccinated regions of genotype space were the virus to connect to other known components.

4.2.3 GA-evolved vaccination strategies

Here we implement a near-canonical GA. Each solution, or vaccination strategy, exist as vector V, whose length equals the number of vaccination strains. V contains the indices of the nodes (strains) to be vaccinated, with values from 1 to network size N.

For a given network, a population of P random solutions is initialized. For up to N_{gen} repetitions, the population is evolved through parent selection based on fitness, crossover, and mutation. Parents are selected through tournament selection with

tournament size T_n. Parents are then recombined via single-point crossover with probability P_c. Indices within each solution are then mutated to a random value from 1 to N according to probability P_m. The best solution at each time step is noted, with the GA exiting before N_{gen} reps if the absolute minimum fitness $F(V, G) = 0$ is found.

4.2.4 Experimental design

Our investigation is three-part: (i) evolving solutions on toy networks, to understand the effects of network structure on solutions, (ii) evolving solution on real-world genotype networks, and (iii) evaluating decay of fitness on a growing network.

In the first part, we constructed the toy networks consisting of a star, lattice, and chain network of size $N = 100$, as well as an Erdős Rényi random network of size $N = 100$, existing as the giant component of a $G(N, p) = G(110, 0.025)$ graph (Figure 1). For 20 repetitions, we ran a GA on each toy network according to the parameters in Table 1. The GA exited when a perfect solution was found ($F(V, G) = 0$) or upon reaching N_{gen} generations. The GA solutions were compared to a distribution of 10^3 random solutions.

In the second part we evolved solutions on a series of real-world influenza A H3N2 genotype networks. These networks were constructed from amino acid sequences of HA observed globally January 2000 through May 2019, sourced from the Influenza Research Database [**?**], in which sequences are represented as nodes and edges exist between sequences differing at one amino acid — indicating a plausible mutation pathway. The real-world networks represent 9 components selected from this network to give a distribution of network sizes from $N = 20$ to $N = 1430$. For both 3 and

Table 4.1: Genetic algorithm parameters

GA Parameter	**Symbol**	**Toy Nets**	**Real Nets**	**Temporal Net**
Population size	P	300	300	200
# vaccine strains	V	3	[3,4]	4
Mutation rate	P_m	1/V	1/V	1/V
Crossover probability	P_c	0.2	0.2	0.2
Max generations	N_{gen}	50	50	20
Tournament size	T_n	2	2	2
Network size (# strains)	N	100	20-1430	$384 \rightarrow 791$
Epidemic threshold	R_0^{crit}	0.5	0.5	0.5
Transcendence	δ	1	[1,2,3]	1

4 vaccination strategies, the GA was run 20 times for each network, for 3 values of transcendence (δ =[1,2,3]) and the parameters found in Table 2. The GA solutions were compared to a distribution of 10^3 random solutions.

In the third part we evaluated changes in fitness as a network grows beyond the time at which a solution was evolved. This simulates vaccination strategies evolved on present strains prior to the emergence of novel strains, at which point fitness may be reduced as the genotype network has grown. Solutions were evolved on a subset of a genotype network of size $N = 791$. The first half of the network to appear ($N = 384$, approximated to the nearest day at which 50% of nodes exist) is used to evolve solutions according to Table 3 across 20 reps. Note that fitness calculations were given knowledge of the full network for accurate genetic distance values. Fitness values were then found for these solutions on the network after 3, 6, and 12 months, as well as for a distribution of 10^3 random networks.

4.2.5 Statistical analysis

To examine the fitness differences between random solutions and GA-derived solutions across the different transcendence values and network sizes, and to analyze the number of function calls required by the different parameter sets, we conducted a series of ANOVAs for each section of our three-part experimental design. For the toy networks, influenza networks, and growing influenza network, we structured our model to examine fitness by group (GA-evolved or random) and the main and interaction effects between network size and transcendence values. To examine the effect of the transcendence value, network size, and their interaction on the number of function calls for the same data sets, we employed an additional three models. In the third part of the study, we examined how random and GA-derived solutions change in fitness over time as the network grows by modeling fitness as a function of group (GA-evolved or random), days after vaccine selection, and their interaction effect. Additionally, we show how the exponential scaling in the number of function calls increases for the size of the network and the number of nodes vaccinated. All analyses were conducted in the R statistical programming language [14].

4.3 Results

The GA was consistently able to derive useful solutions for a combination of different network structures, network sizes and transcendence values. For the toy networks, vaccination strategies selected by the GA showed in a manageable setting how the algorithm took advantage of simple structures to minimize super-critical nodes (Fig-

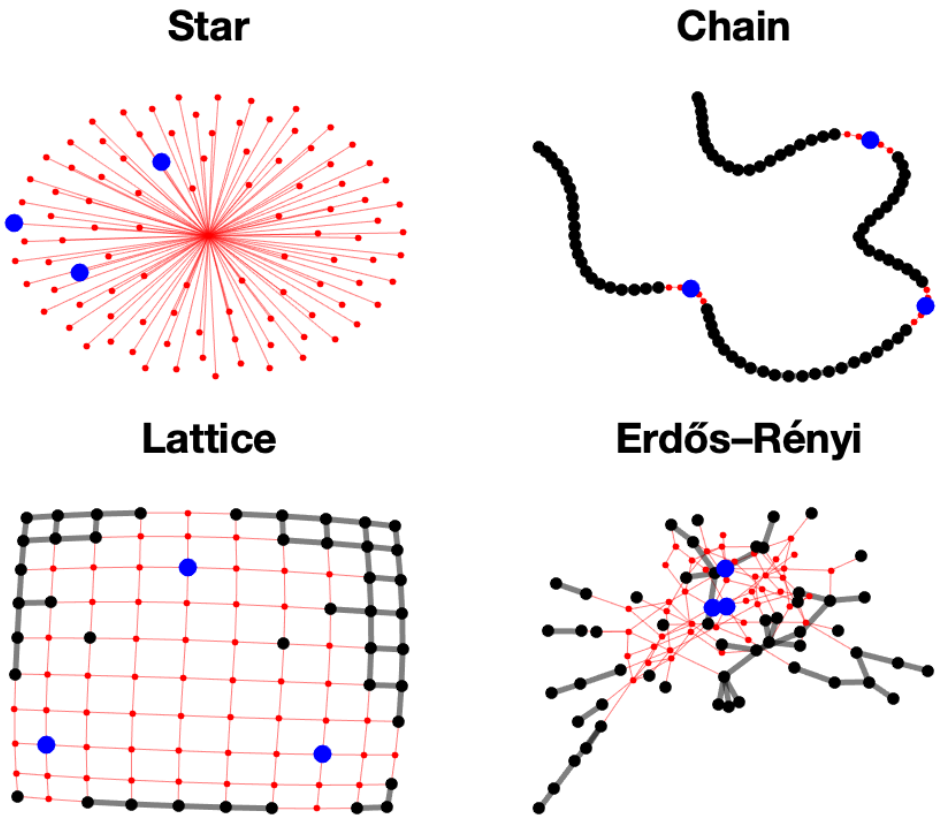


Figure 4.1: Representative vaccination strategy solutions for toy networks. Blue nodes represent strains included in the vaccination. Red nodes indicate nodes that are below the critical threshold for an outbreak. Black nodes are above that threshold.

ure 4.1). There was a difference in the number of function calls required to find a solution between network types, driven by the star network which only needed an average of 200 function evaluations to find a perfect solution (Figure 4.2A). We found in our experimental runs that the GA solutions performed significantly better than the random solutions in terms of fitness ($p < 0.00001$). On average, different networks structures performed differently depending on the transcendence value used ($p < 0.00001$) (Figure 4.2B).

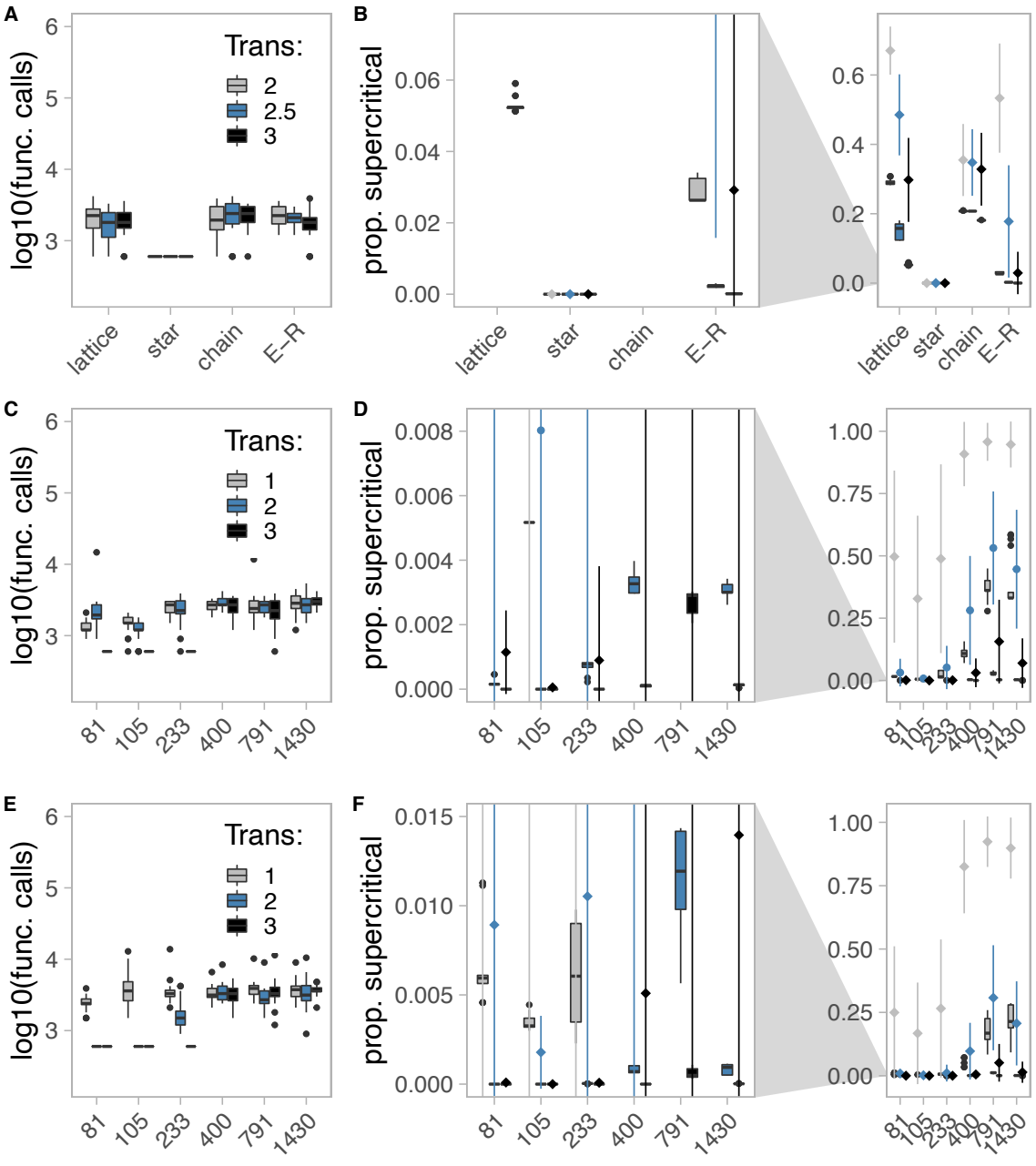


Figure 4.2: (Left most column) The number of function calls (computational effort) on a log scale for each network and transcendence value. (Right most column) The proportion of super-critical nodes to total nodes (fitness) by network for three transcendence values with random solutions with mean shown as diamonds ± 1 standard deviation. Colors grey, blue and black refer to small, medium, and high transcendence values respectively. The middle column is an expanded panel that shows the variance in the smallest distributions of solutions. **A** and **B** refer to the toy networks (lattice, star, chain and Erdős-Rényi) utilizing a vaccination strategy of three vaccines. **C** and **D** refer to the real networks from size 81 to 1430 with a vaccine strategy of three and **E** and **F** refer to a strategy of four on real networks of size 81 to 1430.

For real networks evolved for both 3 and 4 vaccinations, useful strategies were discovered by the GA. A representative example is shown in Figure 4.3. There was a significant difference in the number of function evaluations depending on the network size and transcendence value with of the smaller networks requiring fewer calls ($p<0.00001$) (Figure 4.2C and 4.2E). In terms of fitness, on both 3 and 4 vaccination strategies, the GA performed significantly better than the random solutions ($p<0.00001$) (Figure 4.2D and 4.2F). Again, on average, different networks sizes performed differently depending on the transcendence value used with large networks with low transcendence performing the worst ($p < 0.00001$).

We found that when random and GA solutions were evolved on a portion of a large example genotype network and the network was allowed to grow, the GA solutions performed significantly better than random ones ($p<0.00001$) (Figure 4.4). However, both solutions slowly worsened through time as the network grew ($p<0.00001$). No interaction was observed between time and how the solutions were derived suggesting that both solutions decayed at a similar rate ($p = 0.955$).

The GA was able to find successful solutions on the real networks with linear scaling in the number of function evaluations required to find a workable solution. Figure 4.5 shows the size of the search space for 1, 2, 3 and 4 vaccine strategies on a log_{10} scale. The GA search effort for 4 vaccines, shown as black points, falls in-between the search space of 1 vaccine and 2 vaccines. For the largest real network (N=1430), to search the entire search space for 4 vaccination strategies, $1.74 * 10^{11}$ function evaluations would be required. At the 0.02 seconds it takes for one evaluation, it would take 107.8 years to search the entire network. The GA only performed $3.8 * 10^3$ evaluations on average and found near-perfect ($F = 0$) solutions for transcendence

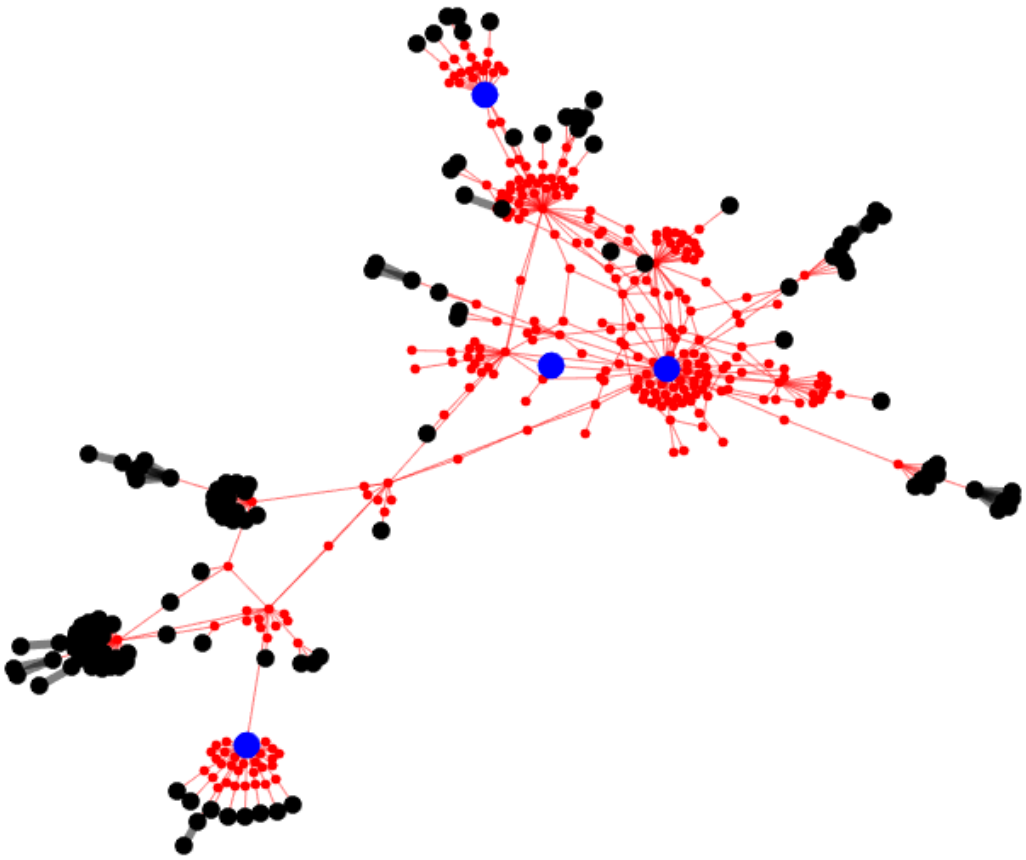


Figure 4.3: Representative vaccination strategy for a moderately sized real flu genotype network ($N = 400$). Blue nodes represent strains included in the vaccination strategy. Red nodes indicate nodes that are below the critical threshold for an outbreak. Black nodes are above that threshold.

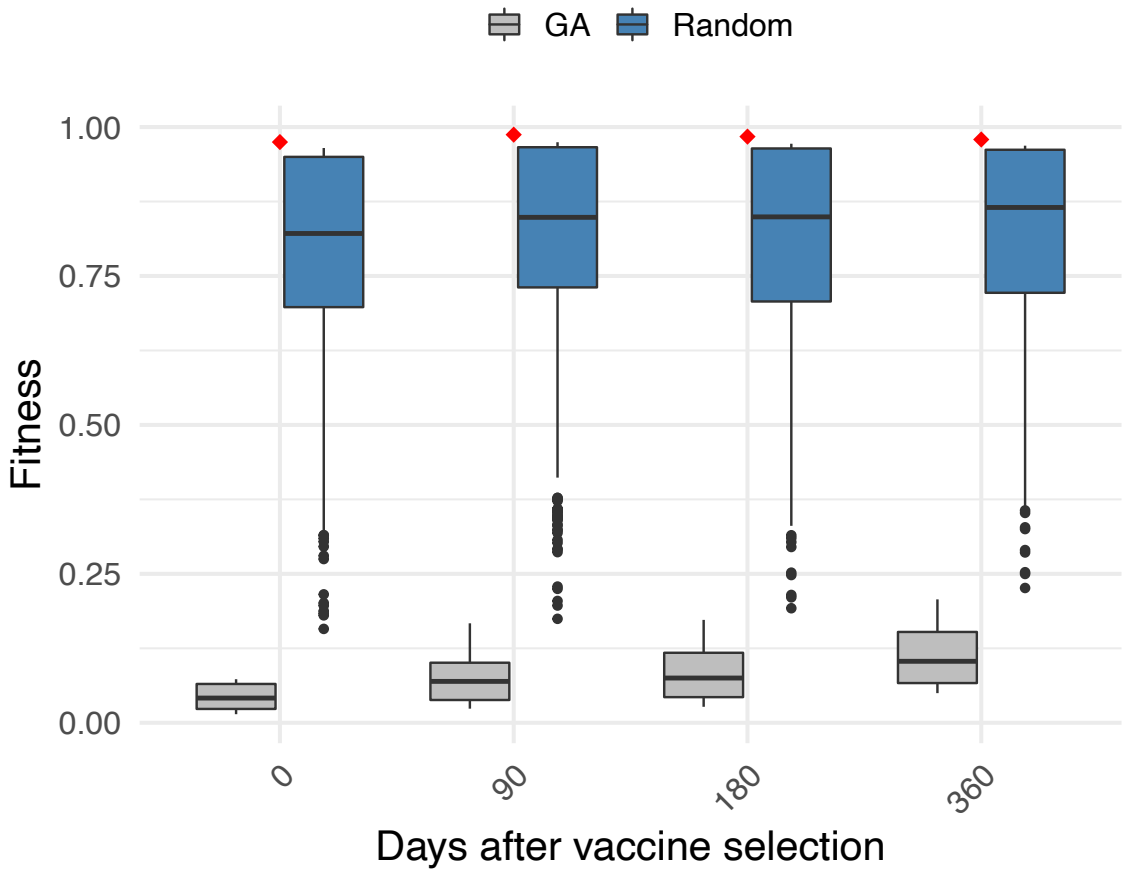


Figure 4.4: Distribution of the fitnesses from GA-evolved and random vaccination strategies on a growing *Influenza A* (H3N2) genotype network, 0, 90, 180, and 360 days after the vaccination strategy was selected.

values of $\delta = [2, 3]$.

4.4 Discussion

4.4.1 Network structure and strategies

Network structure heavily influences both optimal fitness and location of vaccination strains within a genotype network. Fitness measured by the ability of the vaccination strategy to fragment the network into small components allowed for minimization of expected outbreak size (by strain access) in the known genotype space. Thus solutions are rewarded for their ability to not only remove nodes from the network, but to fragment the remaining components. This is seen in the toy networks of Figure 4.1. The chain is broken into 4 nearly if not exactly evenly sized components, minimizing the mean expected outbreak size.

A comparison of the star and the chain indicate the effects of network diameter. Networks of small diameter allow more nodes to fall within the radius of sub-critical influence for a vaccine strain. Although the star and chain are of the same number of nodes, the star's small diameter allows many (if not all) vaccination strategies to provide complete coverage of the known genotype space, indicating that no outbreak would occur. These star-like hubs are found in the influenza networks, whose degree correlates with duplicate samples of a sequence (i.e. greater incidence). Hubs may indicate a particularly virulent or novel strain, yet one whose vaccine would cover a large number of strains, and thus be a target for vaccines. Indeed, hubs were important building blocks for solutions to the influenza networks. However, reducing

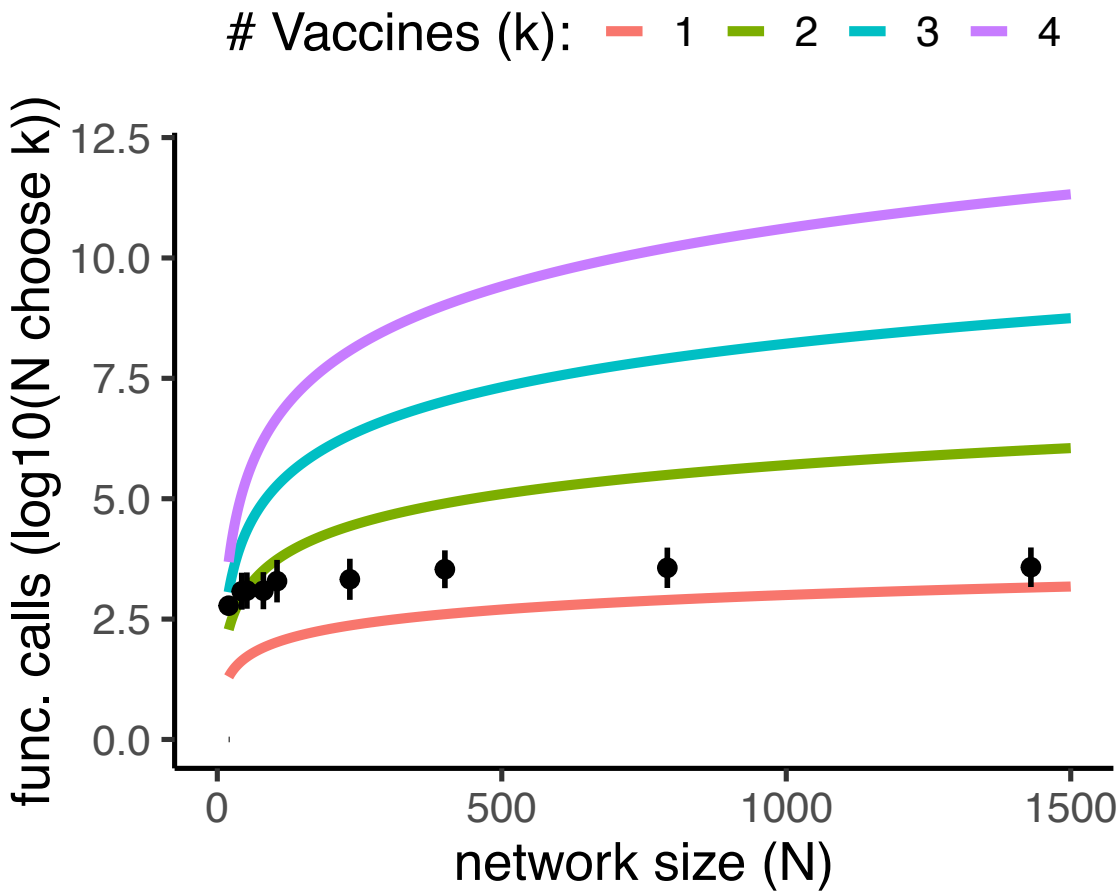


Figure 4.5: Theoretical number of fitness evaluations, $\binom{N}{k}$, for four different numbers of vaccines, k, shown within the size range of real genotype networks, N, (20 to 1430) shown on a log base 10 scale. The mean number of fitness evaluations for 4 vaccines (k=4) required to reach an optimal solution for each real network size are superimposed in black $\pm$ 1 standard error.

total super-critical strains is only one way to reduce outbreak size. Fragmentation of the network into smaller components reduces mean outbreak size, and in larger networks, may only be achievable to significant effect through cooperation between vaccine strains.

The lattice and Erdős-Rényi random graphs demonstrate the cooperation between vaccines, in which only through their combined effect: (i) do some nodes become sub-critical, and (ii) may the network be split at multi-node bridges between large components. The former effect models the case where immunity to multiple strains have a multiplicative or additive effect, or immunity not determined by just the nearest strain. This is an assumption of the model, that the influence from multiple vaccines has a multiplicative effect on immunity, which may be more optimistic than what would be found in transcendence in a real-world application (in comparison to using the maximum immunity, or another interaction between them). The latter effect of cooperation, vaccination at multi-node bridges, implements what may be an important control mechanism on genotype space: blocking evolutionary routes between large or virulent regions of potential protein structures. Deep mutational scanning, which predicts protein stability, could evaluate the effectiveness of targeting these evolutionary bridges by indicating the presence or absence of other pathways between large regions of genotype space [15].

Cooperation between vaccine strain placement is crucial to fragment the network. For instance, a one-strain vaccine strategy on the lattice is optimized with an internal and central placement, while two or more strains must be placed such as to split the network in half, depending on the level of immunity transcendence. In Figure 1 we see the vaccine strains placed to not only make many nodes sub-critical but to isolate

the peripheral regions, reducing mean component size. The Erdős-Rényi network shows a similar strategy: dense central regions remove numerous nodes, while optimal placement fragments the peripheral regions as much as possible.

4.4.2 Real-world vaccination strategies

Vaccination strategies on influenza networks exhibited the same behaviors seen on toy networks. Both 3- and 4-strain vaccination strategies frequently included hubs, while not exclusively using these high degree nodes to fragment the networks. Figure 2 shows a 4-strain strategy on an influenza A (H3N2) HA network of size $N = 400$, that included 3 hubs, while also utilizing a low-degree node to separate the lower-left region component from the upper-right. For transcending immunity levels of $\delta = [1, 2, 3]$ and 3 to 4 vaccine strains, GA-evolved solutions consistently performed better than random solutions ($p < 0.00001$). This demonstrates the superiority of the GA for multistrain vaccine implementation.

The function calls of the GA scaled well with both network size and number of vaccination strains, in addition to tolerating variation in transcendence of immunity (Figure 3, leftmost column). This is in contrast to the $\mathcal{O}\binom{n}{k}$ time complexity of exhaustively searching solutions, visualized in Figure 4.5. This indicates that a simple GA implementation can sufficiently find low-fitness solutions for large search spaces.

4.4.3 Evolved strategies tolerate network growth

GA-evolved vaccine strategies suffered no excess fitness losses relative to random strategies on a growing network. This contradicted our suspicion that random strate-

gies could be more resilient to evolved strategies as novel strains emerged in the genotype network, if their location became more optimal as the network grew. Instead, we see no such advantage in random solutions, as even the best random solutions worsened in time (Figure 4.4). The insignificance of the strategy-by-date interaction (p=0.955) indicates no reduced fitness decay in random strategies. Combined with initial superiority, evolved solutions retain the best fitness values with modest increases for 12 months (Figure 4.4) and beyond. Fitness evaluations beyond 12 months post-solution evolution are not considered, since few strains in the initial portion of the network are likely to be prevalent (thus relevant for vaccine consideration).

Random solutions that improved in time were rare, and it is unlikely to find a random solution with both fitness comparable to GA-evolved solutions and improvements as the network grows. If a random solution were to be found that became better than GA-evolved solutions as the network grew, there would be no justification for its implementation given the unknown future of the structure network. GA-evolved solutions remain superior for coverage of future outbreaks.

4.4.4 Future directions

The fitness function assumes immunity transcends as a logarithmic function of genetic distance between HA sequences of strains, which could be refined by: (i) a more data-driven selection of the transcendence function via HA inhibition assays, such as the experiments that have been conducted on the avian *Influenza A* H5N1 [13] subtype, and (ii) more closely approximating of how multiple acquired immunities combine to affect other strains (*e.g.* multiplicative or additive effects, if not more complex).

More information could be added to the network structure through weighting the

edges by the similarity of the amino acid substitutions between nodes, by using an approach similar to BLOSUM [16]. This could be used to update the transcending immunity between genetically similar strains. Due to local optima observed within the fitness landscape, this GA approach could also be improved by implementing an algorithm that promotes diversity and decreases premature convergence, such as an Age-Layered Population Structure (ALPS) [17], to increase the likelihood that the global optima is found, as well as reducing the need for multiple restarts of the GA.

4.5 Conclusion

Here we identified the features of GA-evolved vaccination strategies on genotype networks and demonstrated their success in reducing expected outbreak size by number of strains. Our approach consistently identified efficacious solutions on a variety of different network structures, sizes, and transcendence values.

The location of vaccination strains within the network greatly influences the overall fitness of the vaccination strategy. A simple GA identifies these optimal vaccine strain selection strategies with considerably less effort than would a brute force search. The GA-evolved solutions were observed to be robust to network growth, resulting from mutations leading to novel strain emergence in real-world viral genotype networks. GA solutions consistently lead to better strategies than random search, across network size, number of vaccine strains, and parameter settings.

We call for investigations that address the following: (i) identification of the viable regions of genotype space, such as through deep mutational scanning, to allow for evolution of vaccination strategies that include future strains, (ii) refinement of the

relationship between genetic similarity of viral strains and the transcendence of immunity, to better inform vaccine coverage, and (iii) evolutionary strategies of vaccine implementation that account for forecasting of active regions of genotype space.

Bibliography

[1] Kendra J Alfson, Gabriella Worwa, Ricardo Carrion, and Anthony Griffiths. Spontaneous Mutation Frequency Notes. *Journal of Virology*, 90(5):2345–2355, 2016.

[2] Yi Guan, Dhanasekaran Vijaykrishna, Justin Bahl, Huachen Zhu, Jia Wang, and Gavin J.D. Smith. The emergence of pandemic influenza viruses. *Protein and Cell*, 1(1):9–13, 2010.

[3] Charles Rupprecht, Ivan Kuzmin, and Francois Meslin. Lyssaviruses and rabies: Current conundrums, concerns, contradictions and controversies. *F1000 Research*, 6(0):1–22, 2017.

[4] Edward A. Belongia, Burney A. Kieke, James G. Donahue, Laura A. Coleman, Stephanie A. Irving, Jennifer K. Meece, Mary Vandermause, Stephen Lindstrom, Paul Gargiullo, and David K. Shay. Influenza vaccine effectiveness in Wisconsin during the 2007-08 season: Comparison of interim and final results. *Vaccine*, 29(38):6558–6563, 2011.

[5] Brendan Flannery, Rebecca J Garten Kondor, Jessie R Chung, Manjusha Gaglani, Michael Reis, Richard K Zimmerman, Mary Patricia Nowalk, Michael L Jackson, Lisa A Jackson, Arnold S Monto, Emily T Martin, Edward A Belongia, Huong Q McLean, Sara S Kim, Lenee Blanton, Krista Kniss, Alicia P Budd, Lynnette Brammer, Thomas J Stark, John R Barnes, David E Wentworth, Alicia M Fry, and Manish Patel. Spread of antigenically drifted influenza A(H3N2) viruses and vaccine effectiveness in the United States during the 2018-2019 season. *The Journal of Infectious Diseases*, 30329:1–8, 2019.

[6] Marie R. Griffin, Arnold S. Monto, Edward A. Belongia, John J. Treanor, Qingxia Chen, Jufu Chen, H. Keipp Talbot, Suzanne E. Ohmit, Laura A. Coleman, Gerry Lofthus, Joshua G. Petrie, Jennifer K. Meece, Caroline Breese Hall, John V. Williams, Paul Gargiullo, La Shondra Berman, and David K. Shay. Effectiveness of non-adjuvanted pandemic influenza A vaccines for preventing pandemic influenza acute respiratory illness visits in 4 U.S. communities. *PLoS ONE*, 6(8):4–10, 2011.

[7] John J. Treanor, H. Keipp Talbot, Suzanne E. Ohmit, Laura A. Coleman, Mark G. Thompson, Po Yung Cheng, Joshua G. Petrie, Geraldine Lofthus, Jennifer K. Meece, John V. Williams, Lashondra Berman, Caroline Breese Hall,

Arnold S. Monto, Marie R. Griffin, Edward Belongia, and David K. Shay. Effectiveness of seasonal influenza vaccines in the United States during a season with circulation of all three vaccine strains. *Clinical Infectious Diseases*, 55(7):951–959, 2012.

[8] Michael L. Jackson, Jessie R. Chung, Lisa A. Jackson, C. Hallie Phillips, Joyce Benoit, Arnold S. Monto, Emily T. Martin, Edward A. Belongia, Huong Q. McLean, Manjusha Gaglani, Kempapura Murthy, Richard Zimmerman, Mary P. Nowalk, Alicia M. Fry, and Brendan Flannery. Influenza vaccine effectiveness in the United States during the 2015–2016 season. *New England Journal of Medicine*, 377(6):534–543, 2017.

[9] D. Steinhauer. Rapid Evolution Of RNA Viruses. *Annual Review of Microbiology*, 41(1):409–433, 1987.

[10] F. Carrat and A. Flahault. Influenza vaccine: The challenge of antigenic drift. *Vaccine*, 25(39-40):6852–6862, 2007.

[11] Scott E. Hensley. Challenges of selecting seasonal influenza vaccine strains for humans with diverse pre-exposure histories. *Current Opinion in Virology*, 8:85–89, 2014.

[12] CDC. Selecting Viruses for the Seasonal Influenza Vaccine, 2018. https://www.cdc.gov/flu/prevent/vaccine-selection.htm.

[13] Thomas Rowe, Robert A. Abernathy, Jean Hu-Primmer, William W. Thompson, Xiuhua Lu, Wilina Lim, Keiji Fukuda, Nancy J. Cox, and Jacqueline M. Katz. Detection of antibody to avian influenza A (H5N1) virus in human serum by using a combination of serologic assays. *Journal of Clinical Microbiology*, 37(4):937–943, 1999.

[14] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019.

[15] Juhye M. Lee, John Huddleston, Michael B. Doud, Kathryn A. Hooper, Nicholas C. Wu, Trevor Bedford, and Jesse D. Bloom. Deep mutational scanning of hemagglutinin helps predict evolutionary fates of human H3N2 influenza variants. *Proceedings of the National Academy of Sciences of the United States of America*, 115(35):E8276–E8285, 2018.

[16] S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, 89(22):10915–10919, 1992.

[17] Gregory S. Hornby. ALPS: The age-layered population structure for reducing the problem of premature convergence. *GECCO 2006 - Genetic and Evolutionary Computation Conference*, 1:815–822, 2006.

Chapter 5

Conclusion

This body of work contributes a further understanding of multistrain disease models and viral genotype networks to the literature of network epidemiology. There is a need to model some diseases not as one pathogen but as multiple interacting pathogens, as is apparent by the multiple epidemic phase transitions and pseudo-chaotic infection progression of the multistrain epidemic model introduced here. Without consideration for multiple strains, accurate forecasting of disease incidence may be limited in time, or even fail to accommodate sufficient infection dynamics.

The challenge of incorporating multiple strains into mathematical epidemic models, with antigenic realism, is addressed with the use of genotype networks. This structure allows for a computationally tractable pairing of genetic information with the framework of existing epidemic models. The informativeness of genotype networks extends beyond epidemic models, with information contained within their structure. The structure alone can tell us where in genotype space a virus or other pathogen is active, where it may be evolving away from, which strains are in wide circulation, and how extensive the roles of mutation and transcending host immunity are in

determining the circulating strains of a pathogen.

Genotype networks are defined in this body of work as one of numerous methods to relate the genetic similarity of multistrain pathogens. A more robust understanding of the antigenic relationship between strains may support alternative definitions of genotype networks, such as defining edges that are weighted by genetic distance, constructing the network based on only the epitope regions of a sequence or the single nucleotide polymorphisms of the sequence, considering the sequences of alternative structures within a pathogen, or redefining the function relating cross-protection to the genotype networks. Given the diversity of multistrain pathogens, the assumptions contained within this body of work that were determined for influenza A viruses may require refinement for applications towards other pathogens. The evolutionary relationship between strains differs from pathogens, and must be considered while constructing and incorporating genotype networks with multistrain epidemic models.

The unique structure and generative processes of genotype networks allows for numerous applications, as explored here. A functional relationship was identified between both age- and degree-weighted preferential attachment, suggesting that a similar relationship may be found in other classes of networks. Methods for multivalent vaccine strain selection were suggested, and with refinement may be considered to predict which vaccination strains are worth exploring.

Altogether, we show the impact of genotype networks on multistrain disease modeling, explore the structure of empirical genotype network structure, and identify applications that include network generative models and vaccine strain selection— highlighting the importance of going beyond the "one disease, one pathogen" paradigm.

Bibliography

[A. J. Kucharski, 2016] A. J. Kucharski, V. Andreasen, J. G. (2016). Capturing the dynamics of pathogens with many strains. *Journal of Mathematical Biology*, 72(1-2):1–24.

[Alfson et al., 2016] Alfson, K. J., Worwa, G., Carrion, R., and Griffiths, A. (2016). Spontaneous Mutation Frequency Notes. *Journal of Virology*, 90(5):2345–2355.

[Allard et al., 2017] Allard, A., Althouse, B. M., Scarpino, S. V., and Hébert-Dufresne, L. (2017). Asymmetric percolation drives a double transition in sexual contact networks. *Proceedings of the National Academy of Sciences*, 114(34):8969–8973.

[Alstott et al., 2014] Alstott, J., Bullmore, E., and Plenz, D. (2014). Powerlaw: A python package for analysis of heavy-tailed distributions. *PLoS ONE*, 9(1):e85777.

[Althouse and Hébert-Dufresne, 2014] Althouse, B. M. and Hébert-Dufresne, L. (2014). Epidemic cycles driven by host behaviour. *Journal of The Royal Society Interface*, 11(99):20140575.

[Althouse et al., 2013] Althouse, B. M., Patterson-Lomba, O., Goerg, G. M., and Hébert-Dufresne, L. (2013). The timing and targeting of treatment in influenza pandemics influences the emergence of resistance in structured populations. *PLOS Comp. Bio.*, 9(2):e1002912.

[Amaral et al., 2000] Amaral, L. A. N., Scala, A., Barthélémy, M., and Stanley, H. E. (2000). Classes of small-world networks. *Proceedings of the National Academy of Sciences*, 97(21):11149–11152.

[Andreasen et al., 1997] Andreasen, V., Lin, J., and Levin, S. A. (1997). The dynamics of cocirculating influenza strains conferring partial cross-immunity. *Journal of Mathematical Biology*, 35(7):825–842.

[Barabási and Albert, 1999] Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509.

[Barabási and Bonabeau, 2003] Barabási, A.-L. and Bonabeau, E. (2003). Scale-free networks. *Scientific American*, 288(5):60–69.

[Barber, 2015] Barber, S., V. J. W. M. (2015). The rate of convergence for approximate bayesian computation. *Electron. J. Statist.*, 9(1):80–105.

[Barrat and Weigt, 2000] Barrat, A. and Weigt, M. (2000). On the properties of small-world network models. *The European Physical Journal B - Condensed Matter and Complex Systems*, 13(3):547–560.

[Bassett and Bullmore, 2006] Bassett, D. S. and Bullmore, E. (2006). Small-world brain networks. *The Neuroscientist*, 12(6):512–523.

[Beaumont, 2010] Beaumont, M. A. (2010). Approximate Bayesian computation in evolution and ecology. *Annual Review of Ecology, Evolution, and Systematics*, 41:379–406.

[Belongia et al., 2011] Belongia, E. A., Kieke, B. A., Donahue, J. G., Coleman, L. A., Irving, S. A., Meece, J. K., Vandermause, M., Lindstrom, S., Gargiullo, P., and Shay, D. K. (2011). Influenza vaccine effectiveness in Wisconsin during the 2007-08 season: Comparison of interim and final results. *Vaccine*, 29(38):6558–6563.

[Blower and Bernoulli, 2004] Blower, S. and Bernoulli, D. (2004). An attempt at a new analysis of the mortality caused by smallpox and of the advantages of inoculation to prevent it. 1766. *Reviews in Medical Virology*, 14(5):275–288.

[Both et al., 1983] Both, G. W., Sleigh, M. J., Cox, N. J., and Kendal, A. P. (1983). Antigenic drift in influenza virus H3 hemagglutinin from 1968 to 1980: multiple evolutionary pathways and sequential amino acid changes at key antigenic sites. *Journal of Virology*, 48(1):52–60.

[Breda et al., 2012] Breda, D., Diekmann, O., de Graaf, W. F., Pugliese, A., and Vermiglio, R. (2012). On the formulation of epidemic models (an appraisal of Kermack and McKendrick). *Journal of Biological Dynamics*, 6 Suppl 2:103–117.

[Bunke and Shearer, 1998] Bunke, H. and Shearer, K. (1998). A graph distance metric based on the maximal common subgraph. *Pattern Recognition Letters*, 19(3):255 – 259.

[Callaway et al., 2000] Callaway, D. S., Newman, M. E. J., Strogatz, S. H., and Watts, D. J. (2000). Network robustness and fragility: Percolation on random graphs. *Physical Review Letters*, 85(25):5468–5471.

[Carrat and Flahault, 2007] Carrat, F. and Flahault, A. (2007). Influenza vaccine: The challenge of antigenic drift. *Vaccine*, 25(39-40):6852–6862.

[Casagrandi et al., 2006] Casagrandi, R., Bolzoni, L., Levin, S. A., and Andreasen, V. (2006). The SIRC model and influenza A. *Mathematical Biosciences*, 200(2):152–169.

[CDC, 2018] CDC (2018). Selecting Viruses for the Seasonal Influenza Vaccine. https://www.cdc.gov/flu/prevent/vaccine-selection.htm.

[CDC, 2019] CDC (2019). Estimates of the incidence of symptomatic influenza by season and age-group, United States, 2010–2016. In *"Key Facts About Influenza"*. https://www.cdc.gov/flu\/about/keyfacts.htm.

[Cohen et al., 2003] Cohen, R., Havlin, S., and Ben-Avraham, D. (2003). Efficient immunization strategies for computer networks and populations. *Physical Review Letters*, 91(24):247901.

[Csilléry et al., 2010] Csilléry, K., Blum, M. G., Gaggiotti, O. E., and François, O. (2010). Approximate Bayesian computation (ABC) in practice. *Trends in Ecology & Evolution*, 25(7):410 – 418.

[de Jong et al., 2007] de Jong, J. C., Smith, D. J., Lapedes, A. S., Donatelli, I., Campitelli, L., Barigazzi, G., Van Reeth, K., Jones, T. C., Rimmelzwaan, G. F., Osterhaus, A. D. M. E., and Fouchier, R. A. M. (2007). Antigenic and genetic evolution of swine influenza A (H3N2) viruses in Europe. *Journal of Virology*, 81(8):4315–4322.

[Diekmann and Heesterbeek, 2000] Diekmann, O. and Heesterbeek, J. (2000). *Mathematical epidemiology of infectious diseases: model building, analysis and interpretation*. Wiley series in mathematical and computational biology. John Wiley and Sons, United States.

[Dietz and Heesterbeek, 2002] Dietz, K. and Heesterbeek, J. (2002). Daniel Bernoulli's epidemiological model revisited. *Mathematical Biosciences*, 180(1):1 – 21.

[Dorogovtsev and Mendes, 2002] Dorogovtsev, S. N. and Mendes, J. F. (2002). Evolution of networks. *Advances in Physics*, 51(4):1079–1187.

[Dorogovtsev and Mendes, 2000] Dorogovtsev, S. N. and Mendes, J. F. F. (2000). Evolution of networks with aging of sites. *Physics Review E*, 62:1842–1845.

[Doyle et al., 2019] Doyle, J. D., Chung, J. R., Kim, S. S., Gaglani, M., Raiyani, C., Zimmerman, R. K., Nowalk, M. P., Jackson, M. L., Jackson, L. A., Monto, A. S., Martin, E. T., Belongia, E. A., McLean, H. Q., Foust, A., Sessions, W., Berman, L., Garten, R. J., Barnes, J. R., Wentworth, D. E., Fry, A. M., Patel, M. M., and Flannery, B. (2019). Interim estimates of 2018-19 seasonal influenza vaccine effectiveness - United States, February 2019. *MMWR Morb Mortal Wkly Rep*, 68(6):135–139.

[Epstein and Price, 2010] Epstein, S. L. and Price, G. E. (2010). Cross-protective immunity to influenza A viruses. *Expert Review of Vaccines*, 9(11):1325–1341.

[Feng and Velasco-Hernández, 1997] Feng, Z. and Velasco-Hernández, J. X. (1997). Competitive exclusion in a vector-host model for the dengue fever. *Journal of Mathematical Biology*, 35(5):523–544.

[Ferguson and Andreasen, 2002] Ferguson, N. and Andreasen, V. (2002). The influence of different forms of cross-protective immunity on the population dynamics of antigenically diverse pathogens. In *Mathematical Approaches for Emerging and Reemerging Infectious Diseases: Models, Methods, and Theory*, pages 157–169, New York, NY. Springer New York.

[Fitch et al., 1991] Fitch, W. M., Leiter, J. M., Li, X. Q., and Palese, P. (1991). Positive Darwinian evolution in human influenza A viruses. *Proceedings of the National Academy of Sciences*, 88(10):4270–4274.

[Flannery et al., 2018] Flannery, B., Chung, J. R., Belongia, E. A., McLean, H. Q., Gaglani, M., Murthy, K., Zimmerman, R. K., Nowalk, M. P., Jackson, M. L., Jackson, L. A., Monto, A. S., Martin, E. T., Foust, A., Sessions, W., Berman, L., Barnes, J. R., Spencer, S., and Fry, A. M. (2018). Interim estimates of 2017-18 seasonal influenza vaccine effectiveness - United States, February 2018. *MMWR Morbidity and Mortality Weekly Report*, 67(6):180–185.

[Flannery et al., 2017] Flannery, B., Chung, J. R., Thaker, S. N., Monto, A. S., Martin, E. T., Belongia, E. A., McLean, H. Q., Gaglani, M., Murthy, K., Zimmerman, R. K., Nowalk, M. P., Jackson, M. L., Jackson, L. A., Foust, A., Sessions, W., Berman, L., Spencer, S., and Fry, A. M. (2017). Interim estimates of 2016-17 seasonal influenza vaccine effectiveness - United States, February 2017. *MMWR Morbidity and Mortality Weekly Report*, 66(6):167–171.

[Flannery et al., 2015] Flannery, B., Clippard, J., Zimmerman, R. K., Nowalk, M. P., Jackson, M. L., Jackson, L. A., Monto, A. S., Petrie, J. G., McLean, H. Q., Belongia, E. A., Gaglani, M., Berman, L., Foust, A., Sessions, W., Thaker, S. N.,

Spencer, S., Fry, A. M., for Disease Control, C., and Prevention (2015). Early estimates of seasonal influenza vaccine effectiveness - United States, January 2015. *MMWR. Morbidity and Mortality Weekly Report*, 64(1):10–15.

[Flannery et al., 2019] Flannery, B., Kondor, R. J. G., Chung, J. R., Gaglani, M., Reis, M., Zimmerman, R. K., Nowalk, M. P., Jackson, M. L., Jackson, L. A., Monto, A. S., Martin, E. T., Belongia, E. A., McLean, H. Q., Kim, S. S., Blanton, L., Kniss, K., Budd, A. P., Brammer, L., Stark, T. J., Barnes, J. R., Wentworth, D. E., Fry, A. M., and Patel, M. (2019). Spread of antigenically drifted influenza A(H3N2) viruses and vaccine effectiveness in the United States during the 2018-2019 season. *The Journal of Infectious Diseases*, 30329:1–8.

[Francis, 1940] Francis, T. (1940). A new type of virus from epidemic influenza. *Science*, 92(2392):405–408.

[Gao et al., 2010] Gao, X., Xiao, B., Tao, D., and Li, X. (2010). A survey of graph edit distance. *Pattern Analysis and Applications*, 13(1):113–129.

[Gershoni et al., 2007] Gershoni, J. M., Roitburd-Berman, A., Siman-Tov, D. D., Freund, N. T., and Weiss, Y. (2007). Epitope mapping. *BioDrugs*, 21(3):145–156.

[Gilbert, 1959] Gilbert, E. N. (1959). Random graphs. *Ann. Math. Statist.*, 30(4):1141–1144.

[Gilchuk et al., 2016] Gilchuk, I., Gilchuk, P., Sapparapu, G., Lampley, R., Singh, V., Kose, N., Blum, D. L., Hughes, L. J., Satheshkumar, P. S., Townsend, M. B., Kondas, A. V., Reed, Z., Weiner, Z., Olson, V. A., Hammarlund, E., Raue, H.-P., Slifka, M. K., Slaughter, J. C., Graham, B. S., Edwards, K. M., Eisenberg, R. J., Cohen, G. H., Joyce, S., and Crowe, James E, J. (2016). Cross-neutralizing and protective human antibody specificities to poxvirus infections. *Cell*, 167(3):684–694.

[Gillespie, 2015] Gillespie, C. S. (2015). Fitting heavy tailed distributions: The poweRlaw package. *Journal of Statistical Software*, 64(2):1–16.

[Girvan et al., 2002] Girvan, M., Callaway, D. S., Newman, M. E., and Strogatz, S. H. (2002). Simple model of epidemics with pathogen mutation. *Physical Review E*, 65(3):031915.

[Gomes et al., 2002] Gomes, M. G. M., Medley, G. F., and Nokes, D. J. (2002). On the determinants of population structure in antigenically diverse pathogens. *Proceedings of the Royal Society B: Biological Sciences*, 269(1488):227–233.

[Grassly and Fraser, 2008] Grassly, N. C. and Fraser, C. (2008). Mathematical models of infectious disease transmission. *Nature Reviews Microbiology*, 6(6):477–487.

[Griffin et al., 2011] Griffin, M. R., Monto, A. S., Belongia, E. A., Treanor, J. J., Chen, Q., Chen, J., Talbot, H. K., Ohmit, S. E., Coleman, L. A., Lofthus, G., Petrie, J. G., Meece, J. K., Hall, C. B., Williams, J. V., Gargiullo, P., Berman, L. S., and Shay, D. K. (2011). Effectiveness of non-adjuvanted pandemic influenza A vaccines for preventing pandemic influenza acute respiratory illness visits in 4 U.S. communities. *PLoS ONE*, 6(8):4–10.

[Guan et al., 2010] Guan, Y., Vijaykrishna, D., Bahl, J., Zhu, H., Wang, J., and Smith, G. J. (2010). The emergence of pandemic influenza viruses. *Protein and Cell*, 1(1):9–13.

[Halstead, 2003] Halstead, S. B. (2003). Neutralization and antibody-dependent enhancement of dengue viruses. *Advances in Virus Research*, 60:421–467.

[Hébert-Dufresne et al., 2013a] Hébert-Dufresne, L., Allard, A., Young, J.-G., and Dubé, L. J. (2013a). Global efficiency of local immunization on complex networks. *Scientific Reports*, 3:2171.

[Hébert-Dufresne et al., 2013b] Hébert-Dufresne, L., Patterson-Lomba, O., Goerg, G. M., and Althouse, B. M. (2013b). Pathogen mutation modeled by competition between site and bond percolation. *Physical Review Letters*, 110(10):108103.

[Henikoff and Henikoff, 1992] Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, 89(22):10915–10919.

[Hensley et al., 2010] Hensley, L. E., Mulangu, S., Asiedu, C., Johnson, J., Honko, A. N., Stanley, D., Fabozzi, G., Nichol, S. T., Ksiazek, T. G., Rollin, P. E., Wahl-Jensen, V., Bailey, M., Jahrling, P. B., Roederer, M., Koup, R. A., and Sullivan, N. J. (2010). Demonstration of cross-protective vaccine immunity against an emerging pathogenic Ebolavirus species. *PLoS Pathogens*, 6(5):e1000904–e1000904.

[Hensley, 2014] Hensley, S. E. (2014). Challenges of selecting seasonal influenza vaccine strains for humans with diverse pre-exposure histories. *Current Opinion in Virology*, 8:85–89.

[Hornby, 2006] Hornby, G. S. (2006). ALPS: The age-layered population structure for reducing the problem of premature convergence. *GECCO 2006 - Genetic and Evolutionary Computation Conference*, 1:815–822.

[Iuliano et al., 2018] Iuliano, A. D., Roguski, K. M., Chang, H. H., Muscatello, D. J., Palekar, R., Tempia, S., Cohen, C., Gran, J. M., Schanzer, D., Cowling, B. J., Wu, P., Kyncl, J., Ang, L. W., Park, M., Redlberger-Fritz, M., Yu, H., Espenhain, L., Krishnan, A., Emukule, G., van Asten, L., Pereira da Silva, S., Aungkulanon, S., Buchholz, U., Widdowson, M.-A., and Bresee, J. S. (2018). Estimates of global seasonal influenza-associated respiratory mortality: a modelling study. *Lancet*, 391(10127):1285–1300.

[Jackson et al., 2017] Jackson, M. L., Chung, J. R., Jackson, L. A., Phillips, C. H., Benoit, J., Monto, A. S., Martin, E. T., Belongia, E. A., McLean, H. Q., Gaglani, M., Murthy, K., Zimmerman, R., Nowalk, M. P., Fry, A. M., and Flannery, B. (2017). Influenza vaccine effectiveness in the United States during the 2015–2016 season. *New England Journal of Medicine*, 377(6):534–543.

[Kamo and Sasaki, 2002] Kamo, M. and Sasaki, A. (2002). The effect of cross-immunity and seasonal forcing in a multi-strain epidemic model. *Physica D: Nonlinear Phenomena*, 165(3):228 – 241.

[Katzelnick et al., 2017] Katzelnick, L. C., Gresh, L., Halloran, M. E., Mercado, J. C., Kuan, G., Gordon, A., Balmaseda, A., and Harris, E. (2017). Antibody-dependent enhancement of severe dengue disease in humans. *Science*, 358(6365):929–932.

[Kermack et al., 1927] Kermack, W. O., McKendrick, A. G., and Walker, G. T. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 115(772):700–721.

[Kermack et al., 1932] Kermack, W. O., McKendrick, A. G., and Walker, G. T. (1932). Contributions to the mathematical theory of epidemics. II. The problem of endemicity. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 138(834):55–83.

[Kermack et al., 1933] Kermack, W. O., McKendrick, A. G., and Walker, G. T. (1933). Contributions to the mathematical theory of epidemics. III. Further studies of the problem of endemicity. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 141(843):94–122.

[Kilbourne, 2006] Kilbourne, E. D. (2006). Influenza pandemics of the 20th century. *Emerging Infectious Diseases*, 12(1):9–14.

[Kish, 1965] Kish, L. (1965). *Survey Sampling*. John Wiley & Sons, New York.

[Koelle et al., 2009] Koelle, K., Kamradt, M., and Pascual, M. (2009). Understanding the dynamics of rapidly evolving pathogens through modeling the tempo of antigenic change: Influenza as a case study. *Epidemics*, 1(2):129 – 137.

[Koutra et al., 2013] Koutra, D., Vogelstein, J. T., and Faloutsos, C. (2013). Deltacon: A principled massive-graph similarity function. *Proceedings of the 2013 SIAM International Conference on Data Mining*, pages 162–170.

[Kucharski et al., 2016] Kucharski, A. J., Andreasen, V., and Gog, J. R. (2016). Capturing the dynamics of pathogens with many strains. *Journal of Mathematical Biology*, 72(1):1–24.

[Kucharski and Gog, 2012] Kucharski, A. J. and Gog, J. R. (2012). Age profile of immunity to influenza: effect of original antigenic sin. *Theoretical Population Biology*, 81(2):102–112.

[Lee et al., 2018] Lee, J. M., Huddleston, J., Doud, M. B., Hooper, K. A., Wu, N. C., Bedford, T., and Bloom, J. D. (2018). Deep mutational scanning of hemagglutinin helps predict evolutionary fates of human H3N2 influenza variants. *Proceedings of the National Academy of Sciences of the United States of America*, 115(35):E8276–E8285.

[Lintusaari et al., 2016] Lintusaari, J., Gutmann, M. U., Dutta, R., Kaski, S., and Corander, J. (2016). Fundamentals and Recent Developments in Approximate Bayesian Computation. *Systematic Biology*, 66(1):e66–e82.

[Lipsitch et al., 2007] Lipsitch, M., Cohen, T., Murray, M., and Levin, B. R. (2007). Antiviral resistance and the control of pandemic influenza. *PLoS Med*, 4(1):e15.

[Marin et al., 2012] Marin, J.-M., Pudlo, P., Robert, C. P., and Ryder, R. J. (2012). Approximate Bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180.

[Meiklejohn and Bruyn, 1949] Meiklejohn, G. and Bruyn, H. B. (1949). Influenza in California during 1947 and 1948. *American Journal of Public Health and the Nation's Health*, 39(1):44–49.

[Michele Berlingerio, 2012] Michele Berlingerio, Danai Koutra, T. E.-R. C. F. (2012). Netsimile: A scalable approach to size-independentnetwork similarity. *ArXiv:1209.2684*.

[Minayev and Ferguson, 2009] Minayev, P. and Ferguson, N. (2009). Improving the realism of deterministic multi-strain models: implications for modelling influenza a. *Journal of The Royal Society Interface*, 6(35):509–518.

[Molinari et al., 2007] Molinari, N.-A. M., Ortega-Sanchez, I. R., Messonnier, M. L., Thompson, W. W., Wortley, P. M., Weintraub, E., and Bridges, C. B. (2007). The annual impact of seasonal influenza in the US: Measuring disease burden and costs. *Vaccine*, 25(27):5086 – 5096.

[Morone and Makse, 2015] Morone, F. and Makse, H. A. (2015). Influence maximization in complex networks through optimal percolation. *Nature*, 524(7563):65–68.

[Nair et al., 2011] Nair, H., Brooks, W. A., Katz, M., Roca, A., Berkley, J. A., Madhi, S. A., Simmerman, J. M., Gordon, A., Sato, M., Howie, S., Krishnan, A., Ope, M., Lindblade, K. A., Carosone-Link, P., Lucero, M., Ochieng, W., Kamimoto, L., Dueger, E., Bhat, N., Vong, S., Theodoratou, E., Chittaganpitch, M., Chimah, O., Balmaseda, A., Buchy, P., Harris, E., Evans, V., Katayose, M., Gaur, B., O'Callaghan-Gordo, C., Goswami, D., Arvelo, W., Venter, M., Briese, T., Tokarz, R., Widdowson, M.-A., Mounts, A. W., Breiman, R. F., Feikin, D. R., Klugman, K. P., Olsen, S. J., Gessner, B. D., Wright, P. F., Rudan, I., Broor, S., Simões, E. A., and Campbell, H. (2011). Global burden of respiratory infections due to seasonal influenza in young children: a systematic review and meta-analysis. *The Lancet*, 378(9807):1917–1930.

[Newman and Watts, 1999a] Newman, M. and Watts, D. (1999a). Renormalization group analysis of the small-world network model. *Physics Letters A*, 263(4):341 – 346.

[Newman et al., 2001] Newman, M. E. J., Strogatz, S. H., and Watts, D. J. (2001). Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 64:026118.

[Newman and Watts, 1999b] Newman, M. E. J. and Watts, D. J. (1999b). Scaling and percolation in the small-world network model. *Physical Review E*, 60:7332–7342.

[Nobusawa and Sato, 2006] Nobusawa, E. and Sato, K. (2006). Comparison of the mutation rates of human influenza A and B viruses. *Journal of Virology*, 80(7):3675–3678.

[Pastor-Satorras et al., 2015] Pastor-Satorras, R., Castellano, C., Van Mieghem, P., and Vespignani, A. (2015). Epidemic processes in complex networks. *Reviews of Modern Physics*, 87:925–979.

[Pastor-Satorras and Vespignani, 2001] Pastor-Satorras, R. and Vespignani, A. (2001). Epidemic spreading in scale-free networks. *Physical Review Letters*, 86:3200–3203.

[Pastor-Satorras and Vespignani, 2002] Pastor-Satorras, R. and Vespignani, A. (2002). Immunization of complex networks. *Physical Review E*, 65(3):036104.

[Patterson-Lomba et al., 2013] Patterson-Lomba, O., Althouse, B. M., Goerg, G. M., and Hébert-Dufresne, L. (2013). Optimizing treatment regimes to hinder antiviral resistance in influenza across time scales. *PloS one*, 8(3):e59529.

[Peeters et al., 2017] Peeters, B., Reemers, S., Dortmans, J., de Vries, E., de Jong, M., van de Zande, S., Rottier, P. J. M., and de Haan, C. A. M. (2017). Genetic versus antigenic differences among highly pathogenic H5N1 avian influenza A viruses: consequences for vaccine strain selection. *Virology*, 503:83–93.

[Polansky et al., 2016] Polansky, L. S., Outin-Blenman, S., and Moen, A. C. (2016). Improved global capacity for influenza surveillance. *Emerging Infectious Diseases*, 22(6):993–1001.

[Putri et al., 2018] Putri, W. C. W. S., Muscatello, D. J., Stockwell, M. S., and Newall, A. T. (2018). Economic burden of seasonal influenza in the United States. *Vaccine*, 36(27):3960–3966.

[R Core Team, 2019] R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[Ratmann et al., 2007] Ratmann, O., Jørgensen, O., Hinkley, T., Stumpf, M., Richardson, S., and Wiuf, C. (2007). Using likelihood-free inference to compare evolutionary dynamics of the protein networks of H. pylori and P. falciparum. *PLoS Computational Biology*, 3(11):e230–e230.

[Robert et al., 2011] Robert, C. P., Cornuet, J.-M., Marin, J.-M., and Pillai, N. S. (2011). Lack of confidence in approximate bayesian computation model choice. *Proceedings of the National Academy of Sciences*, 108(37):15112–15117.

[Roberts et al., 2015] Roberts, M., Andreasen, V., Lloyd, A., and Pellis, L. (2015). Nine challenges for deterministic epidemic models. *Epidemics*, 10:49 – 53.

[Rolfes et al., 2018] Rolfes, M. A., Foppa, I. M., Garg, S., Flannery, B., Brammer, L., Singleton, J. A., Burns, E., Jernigan, D., Olsen, S. J., Bresee, J., and Reed, C. (2018). Annual estimates of the burden of seasonal influenza in the United States: A tool for strengthening influenza surveillance and preparedness. *Influenza and Other Respiratory Viruses*, 12(1):132–137.

[Rowe et al., 1999] Rowe, T., Abernathy, R. A., Hu-Primmer, J., Thompson, W. W., Lu, X., Lim, W., Fukuda, K., Cox, N. J., and Katz, J. M. (1999). Detection of

antibody to avian influenza A (H5N1) virus in human serum by using a combination of serologic assays. *Journal of Clinical Microbiology*, 37(4):937–943.

[Rupprecht et al., 2017] Rupprecht, C., Kuzmin, I., and Meslin, F. (2017). Lyssaviruses and rabies: Current conundrums, concerns, contradictions and controversies. *F1000 Research*, 6(0):1–22.

[Russell et al., 2008a] Russell, C. A., Jones, T. C., Barr, I. G., Cox, N. J., Garten, R. J., Gregory, V., Gust, I. D., Hampson, A. W., Hay, A. J., Hurt, A. C., de Jong, J. C., Kelso, A., Klimov, A. I., Kageyama, T., Komadina, N., Lapedes, A. S., Lin, Y. P., Mosterin, A., Obuchi, M., Odagiri, T., Osterhaus, A. D. M. E., Rimmelzwaan, G. F., Shaw, M. W., Skepner, E., Stohr, K., Tashiro, M., Fouchier, R. A. M., and Smith, D. J. (2008a). The global circulation of seasonal influenza A (H3N2) viruses. *Science*, 320(5874):340–346.

[Russell et al., 2008b] Russell, C. A., Jones, T. C., Barr, I. G., Cox, N. J., Garten, R. J., Gregory, V., Gust, I. D., Hampson, A. W., Hay, A. J., Hurt, A. C., de Jong, J. C., Kelso, A., Klimov, A. I., Kageyama, T., Komadina, N., Lapedes, A. S., Lin, Y. P., Mosterin, A., Obuchi, M., Odagiri, T., Osterhaus, A. D., Rimmelzwaan, G. F., Shaw, M. W., Skepner, E., Stohr, K., Tashiro, M., Fouchier, R. A., and Smith, D. J. (2008b). Influenza vaccine strain selection and recent studies on the global migration of seasonal influenza viruses. *Vaccine*, 26:D31 – D34.

[Sanjuán et al., 2010] Sanjuán, R., Nebot, M. R., Chirico, N., Mansky, L. M., and Belshaw, R. (2010). Viral mutation rates. *Journal of Virology*, 84(19):9733–9748.

[Scarpino and Petri, 2019] Scarpino, S. V. and Petri, G. (2019). On the predictability of infectious disease outbreaks. *Nature Communications*, 10(1):1–8.

[Smith et al., 2004] Smith, D. J., Lapedes, A. S., de Jong, J. C., Bestebroer, T. M., Rimmelzwaan, G. F., Osterhaus, A. D. M. E., and Fouchier, R. A. M. (2004). Mapping the antigenic and genetic evolution of influenza virus. *Science*, 305(5682):371–376.

[Steinhauer, 1987] Steinhauer, D. (1987). Rapid Evolution Of RNA Viruses. *Annual Review of Microbiology*, 41(1):409–433.

[Thomas and Hertz, 2012] Thomas, P. G. and Hertz, T. (2012). Constrained evolution drives limited influenza diversity. *BMC Biology*, 10(1):43.

[Treanor, 2004] Treanor, J. (2004). Influenza vaccine — outmaneuvering antigenic shift and drift. *New England Journal of Medicine*, 350(3):218–220.

[Treanor et al., 2012] Treanor, J. J., Talbot, H. K., Ohmit, S. E., Coleman, L. A., Thompson, M. G., Cheng, P. Y., Petrie, J. G., Lofthus, G., Meece, J. K., Williams, J. V., Berman, L., Breese Hall, C., Monto, A. S., Griffin, M. R., Belongia, E., and Shay, D. K. (2012). Effectiveness of seasonal influenza vaccines in the United States during a season with circulation of all three vaccine strains. *Clinical Infectious Diseases*, 55(7):951–959.

[Uekermann and Sneppen, 2012] Uekermann, F. and Sneppen, K. (2012). Spreading of multiple epidemics with cross immunization. *Physical Review E*, 86(3):036108.

[Van den Hoecke et al., 2015] Van den Hoecke, S., Verhelst, J., Vuylsteke, M., and Saelens, X. (2015). Analysis of the genetic diversity of influenza A viruses using next-generation DNA sequencing. *BMC Genomics*, 16(1):79.

[Wagner, 2014] Wagner, A. (2014). A genotype network reveals homoplastic cycles of convergent evolution in influenza A (H3N2) haemagglutinin. *Proceedings of the Royal Society B: Biological Sciences*, 281(1786):20132763.

[Watts and Strogatz, 1998] Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of small-world networks. *Nature*, 393(6684):440–442.

[Whitley, 1994] Whitley, D. (1994). A genetic algorithm tutorial. *Statistics and Computing*, 4(2):65–85.

[Whitley and Monto, 2006] Whitley, R. J. and Monto, A. S. (2006). Prevention and treatment of influenza in high-risk groups: Children, pregnant women, immunocompromised hosts, and nursing home residents. *The Journal of Infectious Diseases*, 194:S133–S138.

[Wiley and Skehel, 1987] Wiley, D. C. and Skehel, J. J. (1987). The structure and function of the hemagglutinin membrane glycoprotein of influenza virus. *Annual Review of Biochemistry*, 56:365–394.

[Wills and Meyer, 2020] Wills, P. and Meyer, F. G. (2020). Metrics for graph comparison: A practitioner's guide. *PLoS ONE*, 15(2):1–54.

[Zhang et al., 2017] Zhang, Y., Aevermann, B. D., Anderson, T. K., Burke, D. F., Dauphin, G., Gu, Z., He, S., Kumar, S., Larsen, C. N., Lee, A. J., Li, X., Macken, C., Mahaffey, C., Pickett, B. E., Reardon, B., Smith, T., Stewart, L., Suloway, C., Sun, G., Tong, L., Vincent, A. L., Walters, B., Zaremba, S., Zhao, H., Zhou, L., Zmasek, C., Klem, E. B., and Scheuermann, R. H. (2017). Influenza research database: An integrated bioinformatics resource for influenza virus research. *Nucleic Acids Research*, 45(D1):D466–D474.